AI-Driven Adaptive Radiation Treatment Delivery for Head & Neck Cancers

Alexandre Cafaro

université PARIS-SACLAY

AI-Driven Adaptive Radiation Treatment Delivery for Head & Neck Cancers

Radiothérapie adaptative guidée par l'intelligence artificielle pour les cancers ORL

Thèse de doctorat de l'université Paris-Saclay

École doctorale n° 582: cancérologie - biologie - médecine - santé (CBMS)
Spécialité de doctorat: Sciences du Cancer
Graduate School : Life Sciences and Health. Référent : Faculté de médecine

Thèse préparée dans l'unité de recherche **Radiothérapie Moléculaire et Innovation Thérapeutique (INSERM, Institut Gustave Roussy, Université Paris-Saclay)**, sous la direction de **Eric DEUTSCH**, Professeur des universités & praticien hospitalier à l'Université Paris-Saclay, Institut Gustave Roussy, Inserm, la co-direction de **Vincent GREGOIRE**, Professeur des universités & praticien hospitalier au Centre Léon Bérard, la co-direction de **Nikos PARAGIOS**, Professeur à l'Université Paris-Saclay & Président-Directeur Général de Therapanacea, et le co-encadrement de **Vincent LEPETIT**, Professeur à l'École nationale des Ponts et Chaussées.

Thèse soutenue à Paris-Saclay, le 19 Décembre 2024, par

Alexandre CAFARO

Composition du jury

Membres du jury avec voix délibérative

Silke TRIBIUS Professeure - Praticienne hospitalière, Asklepios Hospital St. Georg	Présidente
Hervé DELINGETTE Directeur de recherche, Inria Sophia-Antipolis	Rapporteur & Examinateur
Julia SCHNABEL Professeure, Technical University of Munich	Rapporteure & Examinatrice
Jan-Jakob SONKE Professeur, The Netherlands Cancer Institute	Examinateur

THESE DE DOCTORAT
NNT : 2024UPASL103

Abstract

Head and neck cancer (HNC) is one of the most challenging cancers to treat due to its complex anatomy and significant patient-specific changes during treatment. As the 6^{th} most common cancer worldwide, HNC often has a poor prognosis due to late diagnosis and the lack of reliable predictive markers. Radiation therapy, typically combined with surgery, faces challenges such as inter-observer variability, complex treatment planning, and anatomical changes throughout the treatment process.

Adaptive radiotherapy is essential to maintain precision as the patient's anatomy evolves during treatment. However, current low-invasive imaging methods before each treatment fraction, such as Cone Beam CT (CBCT) and biplanar X-rays, are limited in quality or provide only 2D images, making daily treatment adaptation challenging. This thesis introduces novel deep learning approaches to reconstruct accurate 3D CT images from biplanar X-rays, enabling adaptive radiotherapy that reduces radiation dose, shortens acquisition times, lowers costs, and improves treatment precision.

Reconstructing 3D volumes from biplanar X-rays is inherently challenging due to the limited information provided by only two projections, leading to significant ambiguity in capturing internal structures. To address this, the thesis incorporates anatomical and deformation priors through deep learning, significantly improving reconstruction accuracy despite the very sparse measurements.

The first method, X2Vision, is an unsupervised approach that uses generative models trained on head and neck CT scans to learn the distribution of head and neck anatomies. It optimizes latent vectors to generate 3D volumes that align with both biplanar X-rays and anatomical priors. By leveraging these priors and navigating the anatomical manifold, X2Vision dramatically reduces the ill-posed nature of the reconstruction problem, achieving accurate results even with just two projections.

In radiotherapy, pre-treatment scans such as CT or MRI are typically available and are essential for improving reconstructions by accounting for anatomical changes over time. To make use of this data, we developed XSynthMorph, a method that integrates patient-specific features from pre-acquired planning CT scans. By combining anatomical and deformation priors, XSynthMorph adjusts for changes like weight loss, non-rigid deformations, or tumor regression. This approach enables more robust and personalized

reconstructions, providing an unprecedented level of precision and detail in capturing 3D structures.

We explored the clinical potential of X2Vision and XSynthMorph, with preliminary clinical evaluations demonstrating their effectiveness in patient positioning, structure retrieval, and dosimetry analysis, highlighting their promise for daily adaptive radiotherapy. To bring these methods closer to clinical reality, we developed an initial approach to integrate them into real-world biplanar X-ray systems used in radiotherapy.

In conclusion, this thesis demonstrates the feasibility of adaptive radiotherapy using only biplanar X-rays. By combining generative models, deformation priors, and pre-acquired scans, we have shown that high-quality 3D reconstructions can be achieved with minimal radiation exposure. This work paves the way for daily adaptive radiotherapy, offering a low-invasive, cost-effective solution that enhances precision, reduces radiation exposure, and improves overall treatment efficiency.

Résumé

Le cancer de la tête et du cou (HNC) est l'un des cancers les plus difficiles à traiter en raison de la complexité de son anatomie et des changements significatifs spécifiques à chaque patient au cours du traitement. En tant que 6e cancer le plus fréquent dans le monde, le HNC présente souvent un mauvais pronostic en raison d'un diagnostic tardif et de l'absence de marqueurs prédictifs fiables. La radiothérapie, souvent associée à la chirurgie, est confrontée à des défis tels que la variabilité inter-observateur, la complexité de la planification et les changements anatomiques pendant le traitement.

La radiothérapie adaptative est essentielle pour maintenir la précision à mesure que l'anatomie du patient évolue. Cependant, les méthodes d'imagerie peu invasives actuelles, comme la tomographie conique (CBCT) et les rayons X biplanaires, sont limitées en qualité ou ne fournissent que des images 2D, ce qui complique l'adaptation quotidienne du traitement. Cette thèse propose des approches innovantes basées sur l'apprentissage profond pour reconstruire des images CT 3D précises à partir de rayons X biplanaires, permettant une radiothérapie adaptative qui réduit la dose de radiation, accélère l'acquisition, réduit les coûts et améliore la précision.

La reconstruction de volumes 3D à partir de rayons X biplanaires est difficile en raison des informations limitées de seulement deux projections, ce qui crée une ambiguïté importante dans la capture des structures internes. Pour y remédier, cette thèse intègre des a priori anatomiques et de déformation via l'apprentissage profond, améliorant ainsi considérablement la précision des reconstructions malgré des données limitées.

La première méthode, X2Vision, est une approche non supervisée qui utilise des modèles génératifs entraînés sur des scans CT pour apprendre la distribution des anatomies de la tête et du cou. Elle optimise des vecteurs latents pour générer des volumes 3D alignés avec les rayons X biplanaires et les a priori anatomiques. En utilisant ces a priori et en naviguant dans le domaine anatomique, X2Vision réduit considérablement la nature mal posée du problème de reconstruction, obtenant des résultats précis même avec seulement deux projections.

En radiothérapie, des scans pré-traitement comme le CT ou l'IRM sont souvent disponibles et essentiels pour améliorer les reconstructions en tenant compte des changements anatomiques au fil du temps. Nous avons développé XSynthMorph, une méthode

qui intègre des caractéristiques spécifiques au patient à partir des scans CT préalablement acquis. En combinant des a priori anatomiques et de déformation, XSynthMorph s'adapte aux changements tels que la perte de poids ou les déformations non rigides, permettant des reconstructions plus robustes et personnalisées, avec une précision et un détail sans précédent.

Nous avons exploré le potentiel clinique de X2Vision et XSynthMorph, avec des évaluations cliniques préliminaires montrant leur efficacité dans le positionnement du patient, la recnstruction des structures et l'analyse dosimétrique, soulignant leur potentiel pour la radiothérapie adaptative quotidienne. Pour approcher la réalité clinique, nous avons développé une première approche pour intégrer ces méthodes aux systèmes de rayons X biplanaires utilisés en radiothérapie.

En conclusion, cette thèse démontre la faisabilité de la radiothérapie adaptative utilisant uniquement des rayons X biplanaires. En combinant des modèles génératifs, des a priori de déformation et des scans préalablement acquis, nous avons montré que des reconstructions 3D de haute qualité peuvent être obtenues avec une faible exposition aux radiations. Ce travail ouvre la voie à une radiothérapie adaptative quotidienne, offrant une solution peu invasive, peu coûteuse, et précise.

Acknowledgments

À ma maman

As I reach the culmination of this fantastic journey, I would like to express my deepest gratitude to those who have supported and guided me along the way.
First and foremost, I extend my heartfelt thanks to my four exceptional supervisors:

Prof. Nikos Paragios, Four and a half years ago, you believed in me. Since then, I have grown by your side—not only as an engineer and researcher but also as an individual. Your rigor and vision have been a constant source of inspiration. By following your path, I have learned that we can indeed have an impact and effect change. The countless opportunities for ambitious projects, exchanges, and collaborations you've provided have profoundly shaped me. I have evolved significantly from the student I was a few years ago. With TheraPanacea, you are building an ambitious vision to make the world a better place, and I am proud to be part of it. I will always remember that the sky is the limit and that simplicity is beautiful.

Prof. Eric Deutsch, Your visionary perspective on clinical practice and your approach to integrating future solutions for the best patient outcomes have enlightened me. Being part of the lab at Gustave Roussy was an incredible opportunity to share ideas and engage in stimulating brainstorming sessions.

Prof. Vincent Grégoire, Your vision, rigor, and kindness have been illuminating. Your pragmatic outlook has helped me understand clinical needs, enabling me to better translate engineering concepts into practical applications. Your humor and curiosity about AI have been inspiring. Through your excellence as a doctor, I have gained deeper insight into making a real impact and achieving meaningful integration.

Prof. Vincent Lepetit, Your unwavering support has been invaluable. Your excellence, kindness, and humility have guided me through challenging times, helping me focus on the essentials and find solutions—to see the light amidst the chaos. I have learned so much from your vision and scientific expertise. You have been my mentor always and I am not sure I would have made it without you.

I would also like to thank Charlotte Robert and the team at the Gustave Roussy lab, who welcomed me and guided my clinical research.

My gratitude extends to the members of TheraPanacea who have accompanied and

assisted me throughout this journey. Since arriving as an intern four and a half years ago, they have helped me learn and progress through ambitious projects. Special thanks to Amaury, who has been my main companion during this thesis journey. We shared many memorable moments—brainstorming sessions, lunches, conference travels, and even sports (once!). He has been a model of efficiency, rigor, and intelligence for me. I also want to thank Bastien, with whom I started and made significant progress. Additionally, I am grateful to Ayoub, Sonia, Kumar, Sofiane, Quentin, Paul, Carlos, Despoina, Sami, Olivier, Christophe. Also Jules, and Ethan, whom I had the chance to mentor in turn. Transitioning from receiving to giving is truly inspiring. To the entire fantastic AI and TheraPanacea team, thank you for accompanying me along the way.

My appreciation goes out to my now friends at the Gustave Roussy lab—Julie, Marvin, and Théophraste. We shared and continue to share many moments of laughter and exchanges. I wish you all the best. Also, thank you to Pauline, Léo, Killian, Cathyanne, Ibrahima, and everyone at the lab with whom I shared good times.

I would like to acknowledge the people at Centre Léon Bérard, especially Guillaume and Alexandre, who have always welcomed and guided me.

A special thanks to Prof. Sarah Frisken and Prof. Sandy Wells, who integrated me into the Surgical Planning Laboratory at Brigham and Women's Hospital and Harvard Medical School during my four-month stay. Rarely have I met such kind, excellent, and humble individuals. They nurtured my reflections and broadened my horizons. Thanks also to Tina Kapur and Alexandra Golby for their excellence, kindness and guidance. I thank my now friends Nazim, and Reuben for their guidance, sharing enjoyable NBA games, and fun moments. My time in Boston was truly transformative, opening my eyes to a world of possibilities.

I am also grateful to Sofia Rivera, with whom I worked on breast cancer detection. Your expertise, kindness, and humor helped me understand the needs and how to contribute to better outcomes. Thanks also to Amandine Ruffier and Gabrielle Bielinyte for good shared moments.

To my mother, who has always supported me—you are the most caring person I could hope for. To my father, who instilled in me a love for complexity and humor.

I would like to thank my grandparents from Auvergne, who showed me that with effort, anything is possible. I hope to follow their path of temper, courage, and resilience. The light will always prevail.

I would like to thank my friend El Amine, who has supported me and helped me grow all the way, with kindness and excellence. I maybe wouldn't be on this path without him.

This has been an incredible journey that has allowed me to grow both as a researcher and as a person. It was filled with challenges, intellectual excitement, and complex tasks, balanced between long periods of questioning and pressing deadlines. I have engaged in many brainstorming sessions, discoveries, and shared experiences with amazing people around the me and in the world at conferences. I hope this maturity will guide me in the future. I would be happy to contribute toward making this world a better place.

Contents

List of Figures

List of Tables

Notations and conventions

MISCELLANEOUS

CLB	Centre Léon Bérard	**IGR**	Institut Gustave Roussy
DoF	Degrees of Freedom	**IEC**	International Electrotechnical Commission
HFS	Head-First Supine	**FOV**	Field of View
SID	Source-to-Image Distance	**SOD**	Source-to-Object Distance

MEDICAL

ART	Adaptive Radiation Therapy	**CBCT**	Cone Beam Computed Tomography
CT	Computed Tomography	**IGART**	Image-Guided Adaptive Radiotherapy
DSA	Digital Subtraction Angiograms	**GTV**	Gross Tumor Volume
HNC	Head and Neck Cancer	**HNSCC**	Head and Neck Squamous Cell Carcinoma
HPV	Human Papillomavirus	**HU**	Hounsfield Unit
IGRT	Image-Guided Radiotherapy	**IMRT**	Intensity-Modulated Radiotherapy
IR	Infrared Radiation	**Linac**	Linear Accelerator
MV	Megavoltage	**MRI**	Magnetic Resonance Imaging
OAR	Organ at Risk	**PET**	Positron Emission Tomography
PTV	Planning Target Volume	**QA**	Quality Assurance
RT	Radiation Therapy	**SRS**	Stereotactic Radiosurgery
SPECT	Single Photon Emission Computed Tomography	**3D-CRT**	Three-Dimensional Conformal Radiotherapy

COMPUTER SCIENCE

CNN	Convolutional Neural Network	**CS**	Compressed Sensing
CSGM	Compressed Sensing Using Generative Models	**DCT**	Discrete Cosine Transform
DVF	Deformation Vector Fields	**FBP**	Filtered Back Projection
DRR	Digitally Reconstructed Radiographs	**FID**	Fréchet Inception Distance
GAN	Generative Adversarial Networks	**KL**	Kullback-Leibler Divergence
LASSO	Least Absolute Shrinkage and Selection Operator	**MAE**	Mean Absolute Error
MI	Mutual Information	**MRF**	Markov Random Field
MSE	Mean Squared Error	**MS-SSIM**	Multi-Scale Structural Similarity
NCC	Normalized Cross-Correlation	**PICCS**	Prior Image-Constrained Compressed Sensing
PnP	Perspective-n-Point	**PSNR**	Peak Signal-to-Noise Ratio
SGD	Stochastic Gradient Descent	**SSIM**	Structural Similarity Index Measure
TV	Total Variation	**VAE**	Variational Autoencoder

Chapter 1

Introduction

Contents

1.1 Clinical Context

Head and neck cancer (HNC) represents one of the most intricate challenges in oncology due to the region's complex anatomy and the imperative to preserve vital functions such as speech, swallowing, and sensory perception while administering effective treatment. As the sixth most common cancer worldwide, HNC accounts for approximately 6% of all cancer cases, with over 600,000 new diagnoses annually. The majority are head and neck squamous cell carcinomas (HNSCC), originating from the mucosal surfaces lining the head and neck region.

Often presenting at advanced stages due to nonspecific early symptoms and a lack of reliable screening methods, HNC is associated with a poor prognosis and higher mortality rates. Major risk factors include tobacco use, alcohol consumption, and human papillomavirus (HPV) infection, particularly HPV-16. The intricate anatomy of the head and neck, where vital structures are densely packed in close proximity, adds significant complexity to treatment planning and delivery. Even slight deviations in treatment can lead to substantial morbidity, underscoring the necessity for exceptional precision in therapeutic interventions.

Management of HNC typically involves a multidisciplinary approach that includes surgery, radiation therapy (RT), and chemotherapy. Radiation therapy is a cornerstone of HNC treatment, serving either as a primary modality or in conjunction with other ther-

apies. The objective of RT is to deliver a lethal dose of radiation to the tumor while sparing adjacent healthy tissues to minimize side effects and preserve function. Achieving this balance is particularly challenging in the head and neck region, where the margin for error is minimal due to the presence of critical structures such as the spinal cord, salivary glands, and optic nerves.

1.2 Motivation

Precision in radiation therapy is paramount for HNC patients, whose anatomy can change significantly over the course of treatment due to factors such as tumor regression, weight loss, and tissue swelling. These changes can alter the spatial relationship between the tumor and surrounding healthy tissues, potentially compromising the effectiveness of the treatment and increasing the risk of radiation-induced toxicity.

Traditional RT planning relies on pre-treatment imaging, usually computed tomography (CT) scans, to design treatment plans based on the patient's anatomy at a single point in time. However, this static approach fails to account for anatomical changes occurring during the several weeks of RT, potentially leading to suboptimal dose delivery to the tumor and unintended irradiation of healthy tissues.

Adaptive Radiation Therapy (ART) addresses this limitation by incorporating imaging data acquired throughout the treatment course to adjust the treatment plan as the patient's anatomy evolves. ART has the potential to improve treatment outcomes by maintaining dose conformity to the tumor and reducing toxicity to normal tissues. Implementing ART in clinical practice, however, faces significant challenges, particularly related to the availability of timely, high-quality imaging that does not impose additional radiation burden or extend treatment times.

Current imaging methods used before each treatment fraction, such as Cone-Beam Computed Tomography (CBCT) and biplanar X-rays, have inherent limitations. While CBCT provides volumetric imaging, its image quality is inferior to diagnostic CT, and it requires a full rotational acquisition around the patient. This process leads to additional radiation exposure, increases treatment time, and incurs higher costs. The extended acquisition time adds to patient discomfort and can introduce motion artifacts, while the additional radiation dose contributes to cumulative exposure of organs at risk (OARs), potentially increasing the risk of secondary malignancies.

Biplanar X-rays, in contrast, offer fast, low-dose imaging with minimal radiation exposure and reduced cost. They capture two orthogonal two-dimensional projections, providing essential information for patient positioning based on bony landmarks. However, they lack the volumetric information necessary to accurately assess three-dimensional anatomical changes, limiting their utility in guiding adaptive treatment planning.

Developing methods to reconstruct high-quality 3D images from low-dose, widely available imaging modalities like biplanar X-rays would represent a significant advancement in adaptive radiotherapy. Achieving accurate 3D reconstructions from such sparse data

is inherently challenging due to the limited information and the inherent ambiguity in inferring internal structures from only two projections.

This thesis seeks to overcome these challenges by redefining imaging in adaptive radiotherapy. We envision a system that reconstructs high-fidelity 3D images using only biplanar X-rays, providing CT-quality imaging with minimal radiation exposure, rapid acquisition times, and a cost-effective setup. This innovation promises unprecedented speed, precision, and accessibility in adaptive radiotherapy, potentially transforming clinical practice and improving patient outcomes.

1.3 Contributions

This thesis makes significant contributions to the field of adaptive radiation therapy by developing innovative deep learning methods for accurate 3D reconstruction from biplanar X-rays. By leveraging anatomical and deformation priors, these approaches enable precise 3D imaging for head and neck cancer patients, supporting daily treatment adjustments with reduced radiation exposure and faster acquisition times.

In chapter 2, we provide a comprehensive overview of the challenges in achieving precision in radiotherapy for head and neck cancer. We discuss how traditional methods struggle with anatomical changes like tumor shrinkage and weight loss, leading to misalignment of radiation delivery and increased side effects. We introduce the concept of Image-Guided Adaptive Radiotherapy (IGART), which integrates imaging to dynamically adjust treatment plans. This chapter underscores the critical need for efficient imaging solutions and explores the potential of using limited biplanar X-rays to provide daily 3D estimates of patient anatomy for adaptive radiotherapy with improved precision.

In chapter 3, we address the challenge of 3D reconstruction from biplanar X-rays, recognizing that using only two projections makes the reconstruction problem highly ill-posed due to sparse data. We introduce *X2Vision*, a generative model-based approach that integrates anatomical priors from 3D CT scans to reconstruct head and neck anatomy. By optimizing latent vectors within a deep generative model, *X2Vision* generates a 3D volume that aligns with both the anatomical priors and the input projections, demonstrating improved accuracy over previous methods. Clinical evaluations reveal that *X2Vision* achieves positioning accuracy comparable to CBCT, allowing reliable patient alignment using only biplanar X-rays. Preliminary dosimetric studies further highlight its potential to effectively guide adaptive radiotherapy. This work was presented at MICCAI 2023 (Poster) [Cafaro, 2023c] and ESTRO 2023 (Oral and Poster) [Cafaro, 2023b].

In chapter 4, we enhance the 3D reconstruction accuracy by integrating pre-acquired patient-specific data. We propose *XSynthMorph*, an unsupervised method that combines pre-treatment CT scans with generative models to improve reconstruction from biplanar X-rays. By using a generative model to guide deformations, *XSynthMorph* adapts precisely to each patient's unique anatomy, overcoming limitations of previous methods that struggled with under-constrained deformable alignment. We demonstrate its effectiveness

for adaptive radiotherapy applications, including rigid registration, segmentation, and the deformation of critical organs. This work was presented at WBIR 2024 (Short Oral and Poster) [Cafaro, 2024b] and ESTRO 2024 (Short Oral) [Cafaro, 2024c].

In chapter 5, we adapt our 3D reconstruction frameworks, *X2Vision* and *XSynthMorph*, for integration into real clinical workflows using biplanar X-rays. Translating these methods to real X-ray systems introduces challenges such as partial fields of view, varying imaging geometries, and imaging noise. We adjust our methods to handle these real-world conditions, aiming for seamless integration into clinical radiotherapy workflows. We discuss the obstacles encountered and the innovations introduced to provide precise, personalized adaptive radiotherapy while minimizing radiation exposure.

Finally, in the conclusion, chapter 6, we synthesize our contributions, highlighting the advancements made in 3D reconstruction from biplanar X-rays and their potential to reshape adaptive radiotherapy practices. We discuss future directions for research, such as improving model robustness, integrating additional imaging modalities, and optimizing real-time processing capabilities. Emphasis is placed on the potential impact of these methods on clinical practice, with a focus on enhancing treatment precision, reducing radiation exposure, and ultimately improving patient outcomes through more accessible, adaptive radiotherapy solutions. A patent application encompassing our work has been filed in Europe and the United States [Cafaro, 2024e].

In the Appendix, chapter 7, we present a parallel study focused on 3D cerebral vascular reconstruction from biplanar digital subtraction angiograms (DSAs), offering potential improvements in neurovascular diagnosis and surgical planning. To resolve ambiguities arising from the limited information in 2D projections, we developed a U-Net model for disambiguating backprojected volumes, followed by a MAP refinement with a prior that favors continuity. This approach achieved superior accuracy compared to existing methods. Presented at MICCAI 2024 (Spotlight Oral) [Cafaro, 2024a], this work highlights the promise of biplanar DSAs in enabling accurate, clinically viable 3D vascular reconstructions.

1.4 List of Publications

First Author Publications

[Cafaro, 2023a] A Cafaro, Q Spinat, A Leroy, P Maury, G Beldjoudi, C Robert, E Deutsch, V Grégoire, N Paragios, and V Lepetit. "OC-0443 Full 3D CT reconstruction from partial bi-planar projections using a deep generative model". In: *Radiotherapy and Oncology* (2023). Publisher: Elsevier. Presented at ESTRO 2023.

[Cafaro, 2023b] A Cafaro, Q Spinat, A Leroy, P Maury, G Beldjoudi, C Robert, E Deutsch, V Grégoire, N Paragios, and V Lepetit. "PO-1649 Style-based generative model to reconstruct head and neck 3D CTs". In:

Radiotherapy and Oncology (2023). Publisher: Elsevier. Presented at ESTRO 2023 (cit. on p. 3).

[Cafaro, 2023c] A Cafaro, Q Spinat, A Leroy, P Maury, A Munoz, et al. "X2Vision: 3D CT Reconstruction from Biplanar X-Rays with Deep Structure Prior". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*. Ed. by Hayit Greenspan, Anant Madabhushi, Parvin Mousavi, Septimiu Salcudean, James Duncan, Tanveer Syeda-Mahmood, and Russell Taylor. Lecture Notes in Computer Science. Cham: Springer Nature Switzerland, 2023, pp. 699–709 (cit. on p. 3).

[Cafaro, 2024a] Alexandre Cafaro, Reuben Dorent, Nazim Haouchine, Vincent Lepetit, Nikos Paragios, William M. Wells III, and Sarah Frisken. "Two Projections Suffice for Cerebral Vascular Reconstruction". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*. Ed. by Marius George Linguraru, Qi Dou, Aasa Feragen, Stamatia Giannarou, Ben Glocker, Karim Lekadir, and Julia A. Schnabel. Cham: Springer Nature Switzerland, 2024, pp. 722–731 (cit. on pp. 4, 165).

[Cafaro, 2024b] Alexandre Cafaro, Amaury Leroy, Guillaume Beldjoudi, Pauline Maury, Charlotte Robert, Eric Deutsch, Vincent Grégoire, Vincent Lepetit, and Nikos Paragios. "XSynthMorph: Generative-Guided Deformation for Unsupervised Ill-Posed Volumetric Recovery". In: *International Workshop on Biomedical Image Registration*. Springer. 2024, pp. 19–33 (cit. on p. 4).

[Cafaro, 2024c] Alexandre Cafaro, Amaury Leroy, Guillaume Beldjoudy, Pauline Maury, Alexandre Munoz, Charlotte Robert, Vincent Lepetit, Nikos Paragios, Vincent Grégoire, and Eric Deutsch. "829: 3D CT Reconstruction from biplanar projections with integration of planning CT". In: *Radiotherapy and Oncology* 194 (2024). Publisher: Elsevier. Presented at ESTRO 2024, S3807–S3810 (cit. on p. 4).

[Cafaro, 2024d] Alexandre Cafaro, Amandine Ruffier, Gabriele Bielinyte, Youlia Kirova, Séverine Racadot, Mohamed Benchalal, Jean-Baptiste Clavier, Claire Charra-Brunaud, Marie-Eve Chand-Fouche, Delphine Argo-Leignel, et al. "Abstract PO5-21-03: Cosmetic assessment in the UNICANCER HypoG-01 trial: a deep learning approach". In: *Cancer Research* 84.9_Supplement (2024), PO5–21.

[Cafaro, 2024e] Alexandre Cafaro, Quentin Spinat, Eric Deutsch, Vincent Gregoire, and Nikos Paragios. *3d reconstruction from a limited number of 2d projections*. US Patent App. 18/634,076. Oct. 2024 (cit. on p. 4).

Collaborations

[Buatti, 2024] Jacob S Buatti, Alexandre Cafaro, Sruthi Sivabhaskar, Kristen Duke, Nikos Papanikolaou, Neil Kirby, and Nikos Paragios. "2008: A Generative Adversarial Network for Radiotherapy Dose Predictions of Head and Neck Cancers". In: *Radiotherapy and Oncology* 194 (2024), S3618–S3620.

[Frisken, 2024] SF Frisken, N Haouchine, DD Chlorogiannis, V Gopalakrishnan, A Cafaro, WT Wells, AJ Golby, and R Du. "VESCL: an open source 2D vessel contouring library". In: *International Journal of Computer Assisted Radiology and Surgery* (2024), pp. 1–10.

[Leroy, 2023a] A Leroy, A Cafaro, G Gessain, A Champagnac, V Grégoire, E Deutsch, V Lepetit, and N Paragios. "StructuRegNet: Structure-Guided Multimodal 2D-3D Registration". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*. Ed. by Hayit Greenspan, Anant Madabhushi, Parvin Mousavi, Septimiu Salcudean, James Duncan, Tanveer Syeda-Mahmood, and Russell Taylor. Lecture Notes in Computer Science. Cham: Springer Nature Switzerland. Presented at MICCAI, 2023, pp. 771–780 (cit. on pp. 104, 106).

[Leroy, 2023b] A Leroy, A Cafaro, V Lepetit, N Paragios, E Deutsch, and V Grégoire. "MO-0714 Statistical comparison between GTV and gold standard contour on AI-based registered histopathology". In: *Radiotherapy and Oncology* (2023). Publisher: Elsevier. Presented at ESTRO 2023.

[Leroy, 2023c] A Leroy, A Cafaro, V Lepetit, N Paragios, E Deutsch, and V Grégoire. "OC-0448 Bridging the gap between radiology and histology through AI-driven registration and reconstruction". In: *Radiotherapy and Oncology* (2023). Publisher: Elsevier. Presented at ESTRO 2023.

[Leroy, 2022] A Leroy, M Lerousseau, T Henry, A Cafaro, N Paragios, V Grégoire, and E Deutsch. "End-to-End Multi-Slice-to-Volume Concurrent Registration and Multimodal Generation". In: *Medical Image Computing and Computer Assisted Intervention - MICCAI 2022*. Cham: Springer Nature Switzerland. Presented at MICCAI, 2022, pp. 152–162.

[Leroy, 2024] Amaury Leroy, Alexandre Cafaro, Nazim Benzerdjeb, Anne Champagnac, Grégoire Gessain, Philippe Gorphe, Daphné Morel, Charlotte Robert, Roger Sun, Philippe Zrounba, et al. "1334: Histology to CT transfer for HNC target volume auto-segmentation with deep-learning diffusion models". In: *Radiotherapy and Oncology* 194 (2024), S3047–S3051.

[Spinat, 2024] Quentin Spinat, Despoina Ioannidou, Kumar Shreshtha, Ayoub Oumani, Alexandre Cafaro, Olivier Teboul, and Nikos Paragios. *Method for generating a 3d image of a human body part*. US Patent App. 18/635,312. Oct. 2024.

Chapter 2

Image-Guided Adaptive Radiotherapy for Head and Neck Cancer

In radiotherapy, achieving precision while minimizing side effects is essential—especially in head and neck cancer, where complex anatomy requires accurate targeting to maximize treatment efficacy and preserve vital functions like speech and swallowing.

Traditional radiotherapy relies on a single CT scan, with fractionated doses delivered over multiple sessions. Before each fraction, imaging like Cone-Beam Computed Tomography (CBCT) or biplanar X-rays are used to align the patient with the treatment plan. CBCT offers 3D imaging for precise tissue alignment, while biplanar X-rays mainly enable quick bony alignment. However, as treatment progresses, anatomical changes such as tumor shrinkage, weight loss, or tissue swelling can reduce treatment effectiveness by misaligning radiation and increasing side effects. Adapting the treatment plan to account for these changes is crucial.

Image-Guided Adaptive Radiotherapy (IGART) addresses this challenge by integrating fractional imaging to dynamically adjust plans as anatomy evolves, preserving both accuracy and effectiveness. This chapter explores adaptive strategies for managing these changes, focusing on workflows that use fractionated imaging like CBCT. We also discuss advancements in Adaptive Radiotherapy, emphasizing how imaging quality affects anatomical precision and dose accuracy.

Despite these advancements, the need remains for faster and lower-dose imaging solutions. Could limited biplanar X-rays be leveraged to provide daily 3D estimates of patient anatomy, enabling adaptive radiotherapy with the same precision but with significantly reduced radiation dose, acquisition time, and cost ? This chapter underscores IGART's transformative potential in head and neck cancer treatment, opening the door to more

personalized, precise, and adaptable radiotherapy—and setting the stage for this thesis.

Contents

2.1 Radiotherapy for Head and Neck Cancer

Radiotherapy is a fundamental modality in the treatment of cancer, utilizing ionizing radiation to eliminate malignant cells while preserving healthy tissue. Its application in head and neck cancer presents unique challenges due to the intricate anatomy and the proximity of critical structures. This section provides an overview of radiotherapy and details the standard workflow, emphasizing the importance of precision in treatment planning and delivery for head and neck malignancies.

2.1.1 Introduction to Radiotherapy

Overview of Radiotherapy

Radiotherapy involves the use of high-energy radiation to destroy cancer cells by damaging their DNA, leading to cell death or the inhibition of cell division. The history of radiotherapy dates back to the discovery of X-rays by Wilhelm Conrad Roentgen in 1895 [Röntgen, 1895] and the subsequent discovery of radium by Marie and Pierre Curie [Curie, 1898]. Early in its development, radiotherapy was limited by the lack of precise targeting, which often resulted in significant damage to surrounding healthy tissues.

Today, external beam radiotherapy primarily relies on linear accelerators (Linacs), which generate high-energy megavoltage (MV) photon beams. These beams are directed from multiple angles, enabling precise targeting of the tumor while sparing healthy tissues. This advancement has allowed radiotherapy to evolve into a highly effective, targeted treatment for various cancers.

Over the years, technological advancements have significantly transformed radiotherapy:

- Conventional Radiotherapy: Initially utilized 2D X-ray images for treatment planning, which limited the ability to conform the radiation dose to the exact shape of the tumor.
- Three-Dimensional Conformal Radiotherapy (3D-CRT): Introduced the use of three-dimensional imaging, primarily computed tomography (CT), allowing for better visualization of the tumor and surrounding organs at risk (OARs). This enabled the radiation beams to be shaped more precisely to the tumor volume.
- Intensity-Modulated Radiotherapy (IMRT) Further advanced the precision of dose delivery by modulating the intensity of the radiation beams. IMRT allows for highly conformal dose distributions, sparing normal tissues more effectively than 3D-CRT.
- Image-Guided Radiotherapy (IGRT): Incorporates imaging techniques during treatment delivery to improve accuracy by accounting for patient positioning and anatomical changes.

The overarching goal of radiotherapy remains consistent: to deliver a therapeutic dose to the tumor while minimizing exposure to surrounding healthy tissues. This principle drives continuous innovation in imaging, planning, and delivery techniques to enhance treatment efficacy and reduce side effects.

Radiotherapy in Head and Neck Cancer

Radiotherapy (RT) is a key element in the treatment of head and neck cancers. It can be administered as the primary treatment modality, or in combination with surgery and/or chemotherapy, depending on the stage, location, and specific characteristics of the tumor. For early-stage cancers, radiotherapy alone may be sufficient to achieve good tumor control. In more advanced cases, radiotherapy is often combined with chemotherapy or follows surgical resection to help eliminate any remaining cancer cells and reduce the risk of recurrence.

Advancements in radiotherapy techniques, such as intensity-modulated radiotherapy and image-guided radiotherapy, have improved treatment outcomes by enabling more precise delivery of radiation doses. These methods allow clinicians to shape the radiation dose more accurately to the tumor's geometry, reducing exposure to nearby critical structures.

For the head and neck region, treatment planning must be exceptionally precise due to the complex and densely packed anatomy, as shown in Figure 2.1. Tumors are often located near critical structures like the spinal cord, brainstem, and major blood vessels. Achieving a careful balance between effective tumor control and sparing healthy tissues is essential to minimize the risk of damage to these structures, which could lead to significant impairments in functions like speech, swallowing, and other vital activities.

Treating head and neck cancers is particularly challenging due to several factors:

- Anatomical Complexity: The head and neck region contains a dense network of critical structures, including the spinal cord, brainstem, optic nerves, salivary glands,

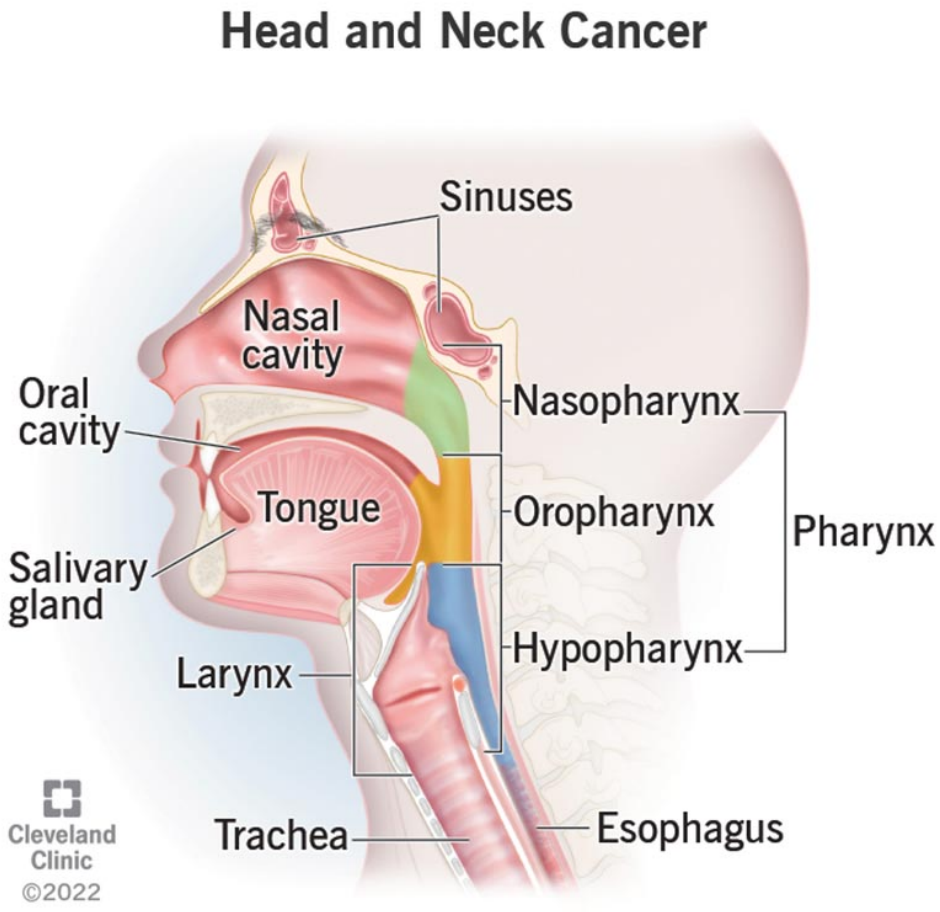

Figure 2.1: Anatomical overview of the head and neck region, highlighting the complex interplay of various structures involved in head and neck cancer. It depicts critical components such as the oral cavity, pharynx, larynx, and surrounding tissues, which are essential for understanding the pathophysiology of cancer in this region. The intricate anatomy presented here underlines the challenges faced in diagnosis and treatment, emphasizing the need for precise imaging and targeted therapies. Illustration from the Cleveland Clinic [Clinic, 2024].

and organs involved in speech and swallowing. These structures are often close to tumor sites, making it difficult to deliver adequate doses to the tumor without risking damage to healthy tissues.

- Functional Preservation: Many of these critical structures are essential for vital functions such as breathing, eating, speaking, and sensory perception. Damage to these areas can greatly impact a patient's quality of life.
- Tumor Heterogeneity: Head and neck tumors vary widely in size, shape, and location, often involving lymphatic spread. This variability necessitates comprehensive coverage of both the primary tumor and regional lymph nodes, adding complexity to treatment planning.
- Interobserver Variability: Due to the complex anatomy and low visibility of certain structures in the head and neck region, even experienced practitioners may differ in delineating target volumes and critical structures. Guidelines, such as those in [Brouwer, 2015; Lefebvre, 2010; Grégoire, 2003], aim to standardize practices, but variability can still occur. This can lead to differences in treatment planning and potentially affect treatment outcomes.

Precision in treatment planning and delivery is then essential in head and neck radiotherapy. Even slight deviations can lead to insufficient tumor dosing or excessive irradiation of normal tissues, resulting in complications such as xerostomia (dry mouth), dysphagia (difficulty swallowing), hearing loss, or radiation-induced neuropathies.

Advancements in radiotherapy techniques, such as IMRT and volumetric modulated arc therapy (VMAT), have improved the ability to conform the radiation dose to complex tumor geometries. However, the need for precise targeting remains critical due to the potential for significant side effects and the desire to preserve normal function.

2.1.2 Radiotherapy Workflow

The radiotherapy process involves several stages, from initial patient assessment to treatment delivery and follow-up. This subsection outlines the standard workflow in radiotherapy, focusing on CT-based radiotherapy and the concept of fractionated treatment delivery.

CT-Based Radiotherapy

The treatment planning process in radiotherapy is crucial for ensuring that the prescribed dose is accurately delivered to the tumor while minimizing exposure to surrounding healthy tissues. The standard workflow consists of the following steps:

1. Patient Positioning Session: A simulation session is conducted to establish and replicate the precise treatment position. Immobilization devices, such as thermoplastic masks for head and neck cases, are employed to maintain consistent positioning throughout the course of treatment, reducing inter-fraction variability.

2. Diagnostic Imaging: A planning CT scan is acquired with the patient in the designated treatment position. This scan provides essential anatomical information, enabling accurate delineation of the tumor and surrounding organs at risk.

3. Multi-Modal Imaging and Fusion: When improved soft-tissue contrast or metabolic information is required, additional imaging modalities such as magnetic resonance imaging (MRI) or positron emission tomography (PET) are acquired. These modalities, which enhance tumor visibility but lack electron density information, are fused with the planning CT to refine anatomical delineations and improve accuracy in defining key structures.

4. Target and Organ at Risk Delineation: In radiotherapy planning, it is crucial not only to target the visible tumor but also to consider potential microscopic disease spread and account for uncertainties. Additionally, protecting surrounding healthy tissues is essential to minimize side effects. To achieve this, several target volumes are defined:

    - Gross Tumor Volume (GTV): The GTV represents the visible or palpable extent of the tumor, as identified by the radiation oncologist based on imaging and clinical examination. It includes only the detectable tumor mass.
    - Clinical Target Volume (CTV): The CTV encompasses the GTV with an additional margin to include areas at risk of microscopic disease spread. This volume ensures that potential regions of subclinical tumor infiltration are also targeted, enhancing comprehensive coverage of cancerous tissue.
    - Planning Target Volume (PTV): To account for setup uncertainties, patient movement, and anatomical variations during treatment, an additional margin is

applied to the CTV to create the PTV. This margin ensures that the prescribed dose consistently covers the entire target area despite slight daily variations in patient positioning or anatomy.

- Organ at Risk (OAR) Delineation: In addition to defining target volumes, critical structures—known as organs at risk (OARs)—are delineated to minimize their radiation exposure. These structures, such as the spinal cord, salivary glands, and optic nerves, are sensitive to radiation and must be carefully spared to reduce the risk of side effects.

5. Treatment Planning: Dosimetrists and medical physicists utilize advanced planning software to develop a treatment plan that specifies the radiation dose distribution, beam arrangements, and modulation techniques (e.g., IMRT) for the Linac. The treatment plan is optimized to maximize PTV coverage while minimizing dose exposure to adjacent OARs.

6. Plan Evaluation and Approval: The treatment plan undergoes comprehensive evaluation for dose conformity, homogeneity, and adherence to established dose constraints for OARs. This final plan is reviewed and formally approved by the radiation oncologist before treatment initiation.

7. Quality Assurance: Prior to commencing treatment, the plan undergoes stringent quality assurance (QA) protocols to verify dose calculation accuracy and validate treatment delivery parameters, ensuring alignment with the initial planning specifications.

In radiotherapy, treatment is typically administered in multiple smaller doses, or fractions, over several weeks. This approach, known as fractionation, provides key biological and clinical benefits. One major advantage is differential repair, where normal tissues can more effectively repair sub-lethal radiation damage compared to cancer cells, thereby reducing the risk of normal tissue complications. Fractionation also allows for repopulation and reoxygenation, as well-oxygenated tumor cells, which are more radiosensitive, can repopulate, making subsequent radiation doses more effective. Additionally, spreading the total radiation dose over multiple sessions helps to minimize acute side effects, improving patient tolerance to the treatment.

For head and neck cancer, treatments are typically delivered five days a week over a period of six to seven weeks, resulting in approximately 30 to 35 fractions. Each fraction delivers a dose to the PTV as defined in the treatment plan.

Reliance on a single CT scan for treatment planning assumes that the patient's anatomy will remain relatively stable throughout the treatment course. However, as we will explore, this assumption often falls short, especially for head and neck cancer patients who commonly experience notable anatomical changes during the treatment period.

2.1.3 Imaging Modalities in Radiotherapy

Imaging is an integral component of radiotherapy, serving multiple purposes from initial diagnosis to treatment planning and delivery. In the context of head and neck cancer, precise imaging is crucial due to the complex anatomy and the need to spare critical structures while delivering an effective dose to the tumor.

Diagnostic Imaging

Diagnostic imaging is the first step in the management of cancer patients, providing essential information on the tumor's location, size, shape, extent, and its relationship to adjacent structures. The primary modalities used in diagnostic imaging for head and neck cancer include:

- Computed Tomography (CT): CT provides detailed cross-sectional images, allowing for clear visualization of bone structures and most soft tissues. CT scans are essential for tumor detection and for delineating organs at risk. In head and neck imaging, contrast agents are often used to enhance visualization of the tumor and lymph nodes.
- Magnetic Resonance Imaging (MRI): MRI provides superior soft-tissue contrast compared to CT, making it valuable for differentiating tumors and surrounding soft tissues.
- Positron Emission Tomography (PET): Often combined with CT (PET/CT), PET detects cellular metabolic activity using radiotracers such as fluorodeoxyglucose (FDG). PET/CT is useful for staging, detecting metastatic disease, and assessing treatment response.
- Ultrasound: This modality is primarily used to evaluate superficial lesions.

The information gathered from diagnostic imaging is critical for staging the disease, formulating a treatment plan, and predicting prognosis. Accurate identification of the tumor and involved lymph nodes ensures that the radiotherapy plan targets all areas of disease while minimizing exposure to healthy tissues.

Figure 2.2: The Elekta Versa HD system with Cone Beam CT (CBCT) imaging. The CBCT provides 3D anatomical visualization by performing a complete gantry rotation around the patient.

Image-Guided Radiotherapy

Image-Guided Radiotherapy uses imaging technologies during each treatment session to improve the precision of radiation delivery. By enabling clinicians to account for patient positioning errors, anatomical changes, and organ motion, IGRT aims to ensure accurate alignment of radiation beams with target volumes as defined in the treatment plan.

The primary goal of IGRT is to achieve accurate alignment, or "rigid fusion", of the patient's current position with the planned treatment position. This alignment is done by matching bony landmarks, fiducial markers, or soft tissue structures. Adjustments can be made with 3 or 6 degrees of freedom, enabling translation and rotation of the treatment table to ensure precise positioning. Before each treatment fraction, 2D or 3D images are acquired and registered with the initial planning CT to check for any misalignments. If needed, adjustments are made to the patient's position to correct both translational and rotational shifts, ensuring accurate radiation delivery.

Imaging systems like CBCT and biplanar X-rays are commonly used in IGRT, each offering unique strengths for patient positioning. Figure 2.4 illustrates a Linac paired with the ExacTrac system and a CBCT imager, both used for precise patient positioning and monitoring in radiotherapy.

Cone Beam Computed Tomography Cone Beam Computed Tomography (CBCT) is a volumetric imaging technique integrated into linear accelerator systems, designed to capture detailed 3D anatomical images for radiotherapy positioning and verification. Using a cone-beam geometry, CBCT typically acquires 200–300 2D X-ray projections during a full

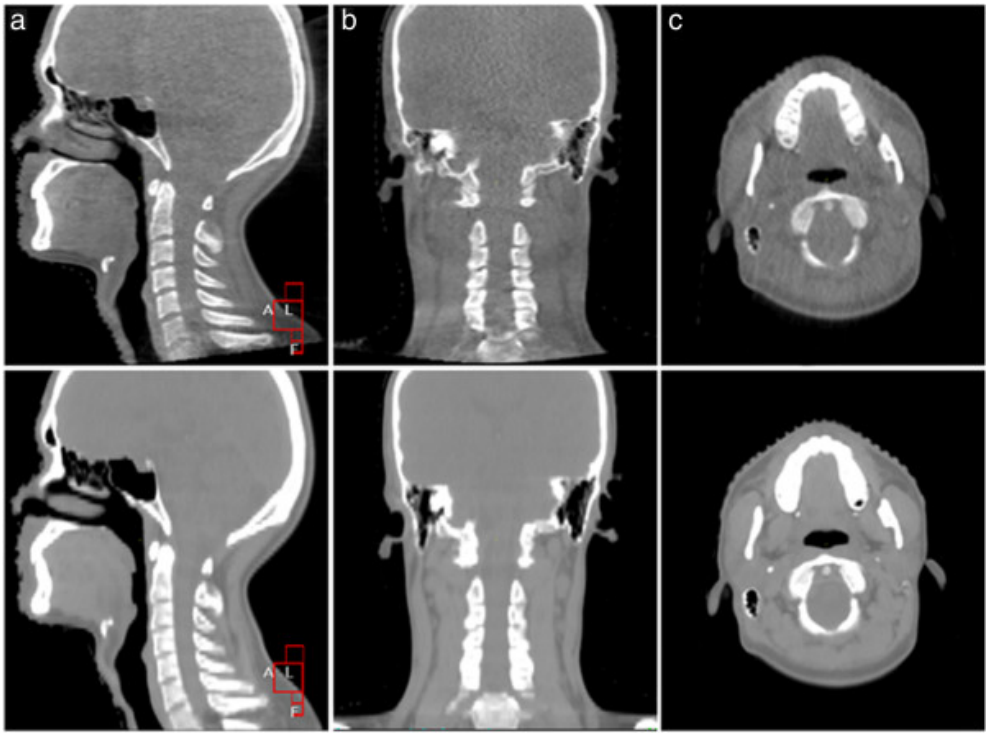

Figure 2.3: Comparison of CBCT and planning CT images for a patient with nasopharyngeal carcinoma. The top row shows cone-beam CT images, while the bottom row displays planning CT images in (a) sagittal, (b) coronal, and (c) transverse planes. Adapted from [Yin, 2013].

360° rotation of the gantry around the patient (as shown in Figure 2.2). These projections are then reconstructed into a 3D image (as shown in Figure 2.3), providing precise visualization of the patient's anatomy in the treatment position. For head and neck imaging, a partial rotation of around 180°–220° is often sufficient, reducing radiation exposure and acquisition time while still capturing essential anatomical details. This 3D view enables accurate alignment of the radiation beams with the target, ensuring that the treatment is accurately directed. Integrated directly into the treatment machine, CBCT allows imaging just before treatment, eliminating the need for patient repositioning. Furthermore, CBCT offers superior soft-tissue contrast compared to traditional portal imaging, improving the visibility of critical anatomical structures needed for precise alignment.

CBCT imaging comes with certain limitations. The acquisition process relies on rotating the C-arm, and the more projections required, the more irradiation and time needed. Additionally, the rotation system is significantly more expensive than fixed X-rays. Depending on the treatment area, acquisition times range from 2 to 4 minutes. Although CBCT generally contributes less to the overall radiation dose than diagnostic CT, it still adds to cumulative exposure, potentially reaching up to 3 Gy over the course of treatment and affecting surrounding healthy organs at risk (OARs) [Spezi, 2012], which may increase the risk of secondary cancers.

Compared to diagnostic CT, CBCT generally suffers from lower image quality due to the inherent limitations of cone-beam geometry, as illustrated in Figure 2.3. Unlike fan-beam CT, CBCT is more prone to scatter radiation, which can introduce artifacts that degrade contrast resolution [Bissonnette, 2008]. This often results in insufficient soft tissue contrast, particularly in complex areas like the head and neck, and challenges in Hounsfield Unit (HU) calibration, which can affect dosimetry accuracy in treatment simulations [Thing, 2016]. To address these limitations, CBCT images are often fused with planning CT to improve anatomical detail and accuracy.

Biplanar X-ray Imaging Biplanar X-ray imaging, used in systems like ExacTrac (Brainlab) [AG, 2024] and the CyberKnife imager (Accuray) [Accuray Incorporated, 2024], captures two orthogonal 2D X-ray images either simultaneously or in rapid succession. Examples of real X-rays from Exactrac are presented in Figure 2.5. For instance, the ExacTrac system combines infrared (IR) tracking with X-ray imaging, using two floor-mounted, non-coplanar X-ray tubes and detectors to capture images from different angles, achieving sub-millimeter accuracy in patient positioning [Jin, 2008]. This accuracy relies on aligning bony landmarks or fiducial markers. The system's real-time monitoring capability allows also for the detection and correction of patient movement during treatment, which is critical for high-precision procedures like stereotactic radiotherapy (SRT). By using only 2 X-rays, biplanar X-ray systems deliver a much lower radiation dose than CBCT, making them suitable for frequent imaging sessions and reducing cumulative exposure in patients who require repeated imaging. Unlike CBCT, biplanar systems do not require rotation; instead, they use two X-ray tubes for immediate image acquisition, providing a

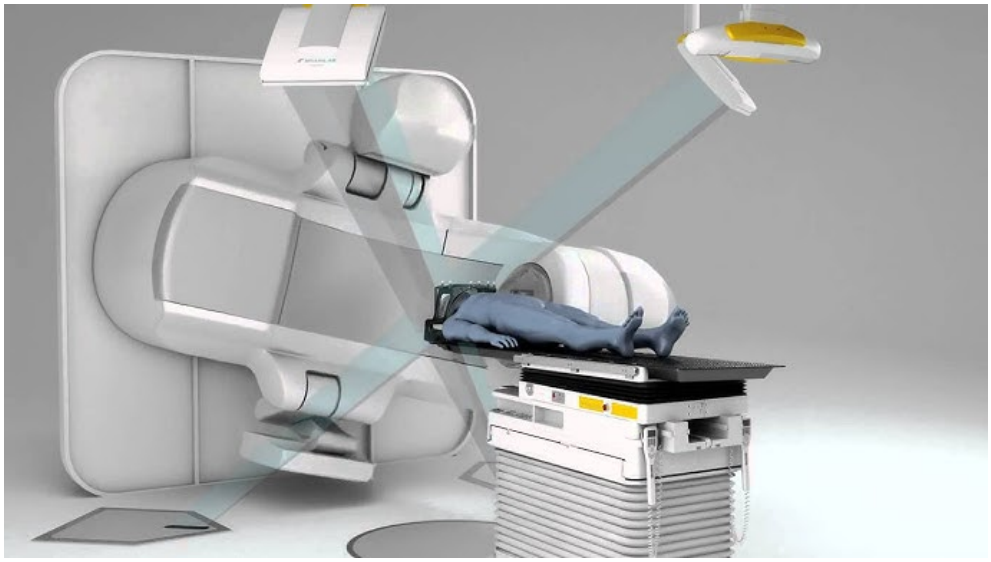

Figure 2.4: A linear accelerator (Linac) equipped with a CBCT imager and the ExacTrac positioning system [AG, 2024] based on biplanar X-rays. The ExacTrac system uses non-coplanar X-rays to capture biplanar 2D images, allowing for fast acquisition and precise positioning based on bony landmarks with minimal radiation exposure.

fast, cost-effective, and low-dose solution.

However, biplanar X-rays capture only planar 2D views, limiting their ability to represent complex 3D anatomical changes. They offer limited soft-tissue contrast and may require fiducial markers in regions lacking clear bony landmarks for accurate tracking. Additionally, their narrower field of view, focused around the PTV, may reduce effectiveness for complete anatomical assessment. A deeper exploration of the ExacTrac system will be presented in Chapter 4.

Both CBCT and biplanar X-ray systems allow rigid fusion for patient positioning, but they serve different purposes. CBCT's volumetric imaging is well-suited for assessing 3D anatomical changes and visualizing soft tissues for precise complex tissue registration. In contrast, biplanar X-ray systems are optimal for precise, low-dose positioning and real-time tracking, especially in treatments requiring high positional accuracy that can rely on bony or visible structural landmarks.

In most protocols, like in radiotherapy centers such as the Centre Léon Bérard (CLB) and the Institut Gustave Roussy (IGR), both systems are often used in conjunction, especially for Volumetric Modulated Arc Therapy (VMAT). Typically, pre-registration is performed using biplanar X-rays. Weekly CBCT scans are then taken to compare registrations. If the CBCT and biplanar X-ray registrations are similar, treatment continues relying solely on biplanar X-rays for the rest of the week. However, if better soft-tissue registration is required, CBCT is then acquired to refine the positioning.

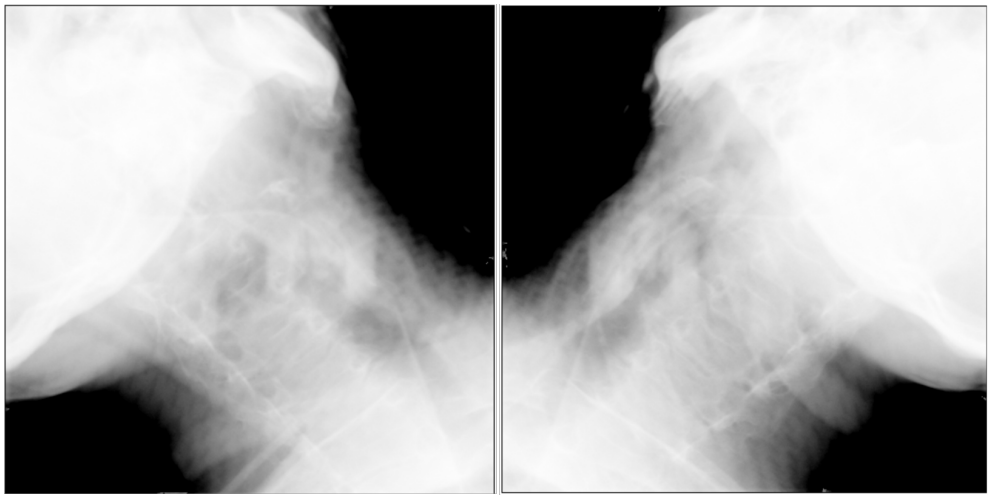

Figure 2.5: Examples of real biplanar X-rays from the ExacTrac system, which uses two non-coplanar X-ray tubes for orthogonal 2D imaging. This approach enables precise patient positioning by aligning bony landmarks or fiducial markers, with low radiation dose suitable for frequent imaging. However, biplanar X-rays provide very limited 3D anatomical detail and soft-tissue contrast.

MRI Also, MRI provides high soft-tissue contrast without ionizing radiation, making it ideal for radiotherapy, particularly in complex regions like the head and neck. Recent advances have integrated MRI with linear accelerators, creating MR-Linac systems like the Elekta Unity [Elekta, 2024] and ViewRay MRIdian [ViewRay Incorporated, 2024]. These systems can enable real-time imaging during radiation delivery, allowing for adaptive radiotherapy based on dynamic anatomical changes, supporting more accurate tumor targeting. MRI offers benefits such as enhanced soft-tissue delineation and no additional radiation dose.

However, challenges include technical integration issues, lack of electron density information for dose calculations, and high cost.
Nevertheless, classical IGRT alone may not fully address the anatomical changes that occur during the course of treatment in head and neck cancer patients. While IGRT improves the precision of radiation delivery by correcting for daily setup errors and minor anatomical variations, significant changes such as tumor shrinkage, weight loss, and tissue swelling can alter the patient's anatomy beyond the capabilities of IGRT to compensate. These substantial variations may necessitate adjustments to the treatment plan itself—a concept central to adaptive radiotherapy.

2.2 Adaptive Radiotherapy

Adaptive Radiotherapy (ART) is a significant advancement in radiation oncology, improving treatment efficacy and reducing toxicity by adapting to the anatomical and physiological changes that occur during the course of radiotherapy. This approach is especially valuable in head and neck cancer, where frequent anatomical changes, such as tumor shrinkage and weight loss, can markedly affect treatment accuracy.

The concept of ART, introduced by Di Yan [Yan, 1997], centers on using an imaging feedback loop to account for patient-specific anatomical variations in treatment planning. A comprehensive ART system involves four main components [Sonke, 2019]: (i) assessing the treatment dose to ensure accuracy, (ii) identifying and evaluating treatment variations, (iii) making informed decisions on treatment modifications, and (iv) implementing adaptive modifications to the treatment plan. By continuously adjusting the treatment plan to account for anatomical changes, ART keeps the radiation dose precisely targeted to the tumor, ensuring optimal dose delivery, minimizing exposure to healthy tissues, and reducing the likelihood of side effects.

2.2.1 The Need for Adaptive Radiotherapy

In head and neck cancer, patients often undergo notable anatomical changes during radiotherapy due to factors like tumor shrinkage, weight loss, edema, and variations in positioning. These changes can shift the spatial relationships between the tumor, OARs, and surrounding anatomy, leading to discrepancies between the planned and delivered dose distributions.
Anatomical changes in head and neck during radiotherapy include :

- Tumor Shrinkage : radiation therapy can effectively reduce tumor size over time as cancer cells are destroyed. As the tumor shrinks, its position relative to surrounding tissues and OARs may change. This shift can result in underdosing of the tumor if it moves out of the high-dose region defined in the initial treatment plan. Barker et al. quantified volumetric and geometric changes during fractionated radiotherapy for head and neck cancer, demonstrating significant tumor regression in many patients [Barker, 2004].

- Weight Loss and Changes in Body Contour : patients undergoing head and neck radiotherapy frequently experience weight loss due to side effects like mucositis, dysphagia, and altered taste sensation. Weight loss leads to a reduction in subcutaneous fat and muscle mass, altering the patient's external contour and internal anatomy. These changes can affect the fit of immobilization devices and result in positioning errors. Navran et al. reported that significant weight loss during treatment can impact dose distribution and increase the risk of toxicity [Navran, 2019].

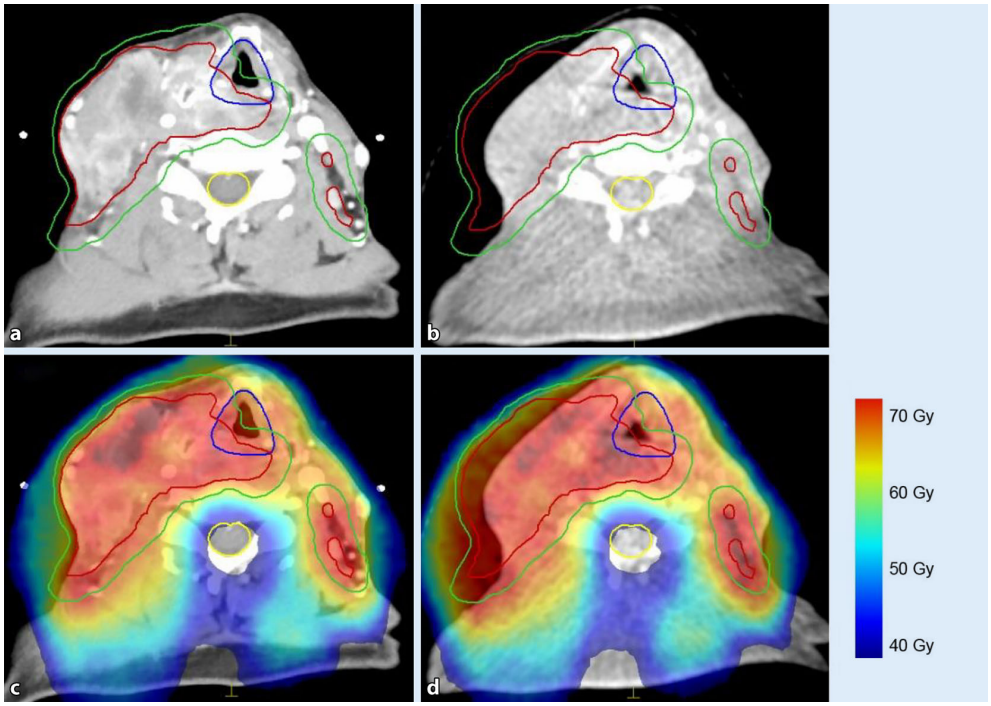

Figure 2.6: Head and neck anatomy changes during radiotherapy necessitating adaptive strategies. Shown here is a planning CT scan (a) with dose overlay (c) and the same patient in week 5 of treatment (b, d). Despite evident regression of treated lymph nodes (b), dose coverage remains unchanged and hence too spread without adaptation in this photon-based rotational arc radiotherapy (d). Key anatomical structures: GTV (red contour), PTV (green contour), larynx (blue contour), and spinal canal (yellow contour) [Herrmann, 2015].

- Edema and Tissue Swelling : radiation-induced inflammation can cause swelling of tissues (edema), leading to changes in the size and shape of both the tumor and adjacent normal tissues. Swelling can alter dose distribution by changing the distance between the radiation source and the target or by shifting OARs closer to the high-dose region. This can increase the risk of side effects and complicate dose delivery.

- Patient Positioning Variations : despite the use of immobilization devices, slight variations in patient positioning can occur between treatment sessions due to discomfort, changes in neck flexibility, or differences in setup procedures. These positional variations can lead to misalignment between the radiation beams and the target volumes, potentially resulting in underdosing of the tumor or overexposure of OARs [Castadot, 2010].

These anatomical shifts can lead to two main issues: underdosing the tumor and overexposing OARs. As the tumor shrinks or shifts, it may fall outside the high-dose region, reducing the intended radiation effect. Similarly, if OARs move into the radiation field, they may receive higher doses than planned, raising the risk of toxicity and side effects like xerostomia or dysphagia. Figure 2.6 represents these anatomical changes and induced effects without adaptation.

ART adapts the treatment plan during therapy to match the patient's current anatomy, ensuring the tumor stays within the high-dose region despite changes in size or position. This reduces toxicity by limiting radiation to shifted OARs and enhances treatment efficacy by accurately targeting the tumor, improving control and minimizing side effects.

2.2.2 Clinical Outcomes of Adaptive Radiotherapy

In head and neck cancer, ART provides substantial clinical advantages. By continuously adapting treatment plans in response to anatomical changes, ART ensures precise dose delivery to the tumor, leading to improved local disease control and heightened treatment efficacy [Castadot, 2010; Kranen, 2013]. Research suggests that ART achieves superior dosimetric outcomes, offering more consistent tumor coverage while limiting exposure to healthy tissues, which translates into enhanced clinical results [Hussein, 2018; Capelle, 2012].

The adaptability of ART in shaping the radiation field reduces exposure to surrounding OARs, lowering toxicity and decreasing the likelihood of side effects such as xerostomia and dysphagia, prevalent among head and neck cancer patients [Schwartz, 2013].

This decrease in adverse effects positively influences patient quality of life, helping to preserve essential functions like swallowing and speech. Consequently, ART contributes to improved long-term functional outcomes and heightened patient satisfaction [Navran, 2019; Heijkoop, 2014].

However, ART is not yet widely used in clinical settings and remains primarily a focus of research due to several barriers [Sonke, 2019]. These include the need for advanced

imaging and computational resources, increased treatment planning time, and workflow complexities that require specialized training for clinicians. Additionally, frequent anatomical monitoring can be resource-intensive, which limits its feasibility in routine clinical practice. We will present the envisioned strategies to implement ART.

2.2.3 Methods of Adaptive Radiotherapy

Adaptive Radiotherapy encompasses various strategies to adjust radiation treatment plans in response to anatomical changes during therapy. These methods are categorized by the timing of adaptation, the frequency of plan modifications, and the specific triggers for adjustments. ART relies on imaging during each fraction to monitor changes in patient anatomy and adapt the treatment accordingly. The goal is to ensure that the cumulative dose—the total radiation dose delivered over the entire treatment course—aligns closely with the planned dose distribution, thereby protecting organs at risk and maintaining effective tumor coverage.

Figure 2.7 illustrates different ART typologies for head and neck cancer, with approaches that vary in temporal resolution. These range from single adaptations to more advanced methods that involve continuous dose accumulation and daily plan modifications, ensuring that the cumulative dose respects the intended treatment plan despite anatomical changes [Sonke, 2019].

ART methods are generally classified as either offline or online [Heukelom, 2019]:

- Offline adaptive methods involve treatment plan adjustments made between sessions based on prior imaging, usually using data from weekly or periodic scans (e.g., CBCT or CT). This allows the treatment plan to be updated without interrupting daily therapy sessions, providing flexibility in adapting to anatomical changes over time.
- Online adaptive methods enable real-time plan modifications directly before or even during treatment sessions. These methods rely on in-session imaging and rapid replanning, allowing for immediate adaptations to account for day-to-day anatomical changes, thereby enhancing the accuracy of dose delivery.

Fixed-Interval Adaptation

In Fixed-Interval Adaptation (Figure 2.7A), the treatment plan is adjusted at predetermined time points during the therapy course, often at mid-treatment. The initial plan is recalculated or superimposed on a new image acquired at the scheduled interval, typically a repeat CT scan. If dose constraints are not met due to anatomical changes, a single adaptation is performed based on the updated anatomy [Heukelom, 2019].

This approach is computationally efficient and integrates smoothly into clinical workflows. It is particularly useful in scenarios where anatomical changes are expected to occur gradually and can be addressed with periodic adjustments. However, it may not promptly respond to unexpected significant alterations that occur outside the scheduled intervals.

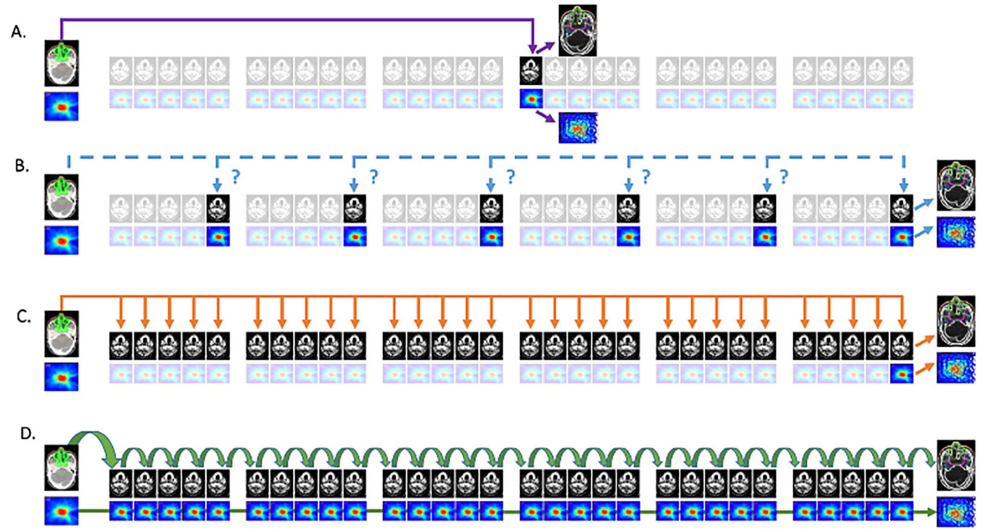

Figure 2.7: Typologies of ART implementation: (A) Fixed-Interval Adaptation, (B) Triggered Adaptation, (C) Serial Adaptation, and (D) Cascaded Adaptation [Heukelom, 2019].

Triggered Adaptation

Triggered adaptation (Figure 2.7B), involves modifying the treatment plan when specific clinical criteria or thresholds are met during the course of treatment. These triggers are identified through routine imaging or clinical assessments and may include significant tumor shrinkage, substantial weight loss causing changes in body contour, shifts or deformations in OARs, or increased setup deviations due to changes in mask fit [Sonke, 2019].

When these predefined thresholds are exceeded, an adaptive plan is created to reflect the updated anatomical situation, optimizing resource utilization by adapting the treatment only when necessary and tailoring the approach to individual patient changes. For example, Møller et al. [Møller, 2016] implemented a triggered adaptation protocol in non-small cell lung cancer patients, where daily CBCT scans for soft-tissue matching were systematically evaluated, and plans were adapted if residual uncertainties in tumor or lymph node positions exceeded certain thresholds over three consecutive fractions.

Triggered adaptation can also be combined with fixed-interval adaptation to provide greater flexibility. Schwartz et al. used a protocol that included a single fixed-interval adaptation, with additional triggered adaptations performed as necessary based on observed anatomical changes [Schwartz, 2013].

Serial or Sequential Adaptation

In Serial Adaptation (Figure 2.7C), also known as "one-to-many" or "sequential" adaptation, the treatment plan is frequently updated—often weekly—based on new imaging data, such as CBCT or MRI. Each update registers the latest images to the original planning images, focusing solely on the current anatomy.

However, cumulative dose accumulation is not performed; the method does not consider the dose delivered in previous fractions [Heukelom, 2019]. As a result, changes in OARs and target volumes during therapy are not factored into dose calculations, which may affect the precision of the delivered dose.

Online Replanning

Online Replanning modifies the treatment plan in real-time, immediately before or during each session. Imaging is performed just prior to treatment, capturing the patient's current anatomy, and the plan is adjusted "on-the-fly" to account for any observed anatomical changes, ensuring accurate dose delivery to the target while sparing healthy tissues [Sonke, 2019]. This method addresses day-to-day anatomical variations, correcting both systematic and random errors, but focuses only on the current session, without considering the cumulative dose from previous treatments.

While online replanning enhances per-session accuracy, it treats each fraction independently, which may limit cumulative dose optimization for the tumor and OARs. Additionally, rapid imaging, contouring, and plan optimization are required for maintaining treatment efficiency, and advanced technologies like MRI-guided radiotherapy systems—such as the Elekta Unity[Elekta, 2024]—have enabled high-quality, real-time imaging and adaptation within a single session [Bohoudi, 2017]. Automation in segmentation and planning further streamlines the process, reducing clinician workload [Hussein, 2018]. Despite logistical challenges, online replanning offers a highly individualized ART approach, with potential for improved treatment accuracy and patient outcomes.

Cascaded or Iterative Adaptation

Cascaded Adaptation (Figure 2.7D), also known as "iterative" adaptation, represents the most comprehensive approach to ART. Unlike online replanning, cascaded adaptation not only adjusts the treatment plan based on the current anatomy but also incorporates the cumulative dose delivered in all previous treatment fractions into the planning process [Sonke, 2019; Heukelom, 2019].

In this method:

- Daily Imaging and Deformable Registration: Imaging is performed daily to capture anatomical changes for each treatment fraction. Deformable image registration is used to map these daily images to a common reference frame.

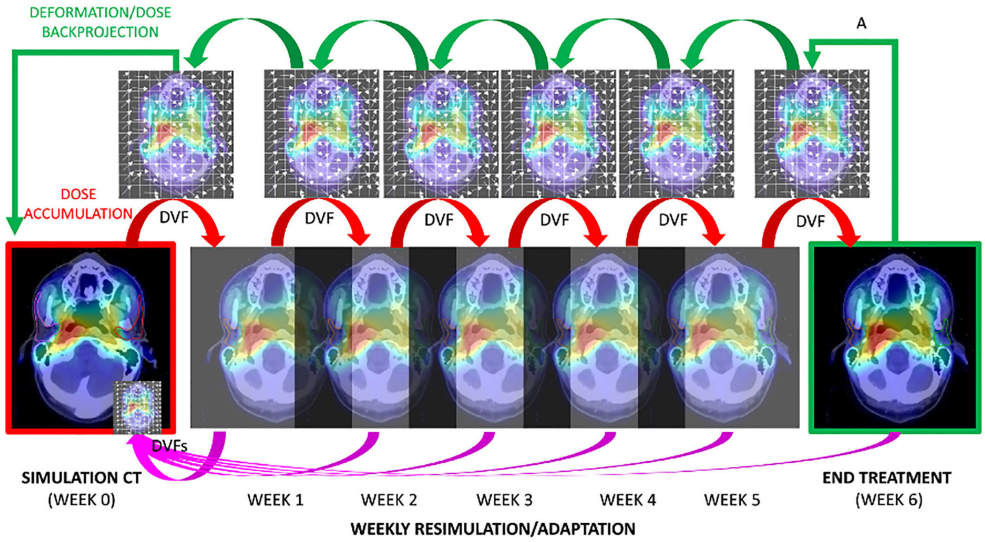

Figure 2.8: Weekly cascade registrations using deformable image registration onto fractionated CBCTs and dose accumulation through deformation vector fields (DVFs), showing progressive parotid gland volume changes and dose tracking across therapy sessions [Heukelom, 2019].

- Concatenation of Deformation Vector Fields (DVFs): DVFs from each fraction are concatenated to track cumulative anatomical deformations over time, creating a comprehensive model of how the patient's anatomy has changed throughout the treatment course.
- Concurrent Dose Accumulation: The radiation dose delivered in each fraction is accumulated using the concatenated DVFs, allowing for an accurate assessment of the cumulative dose distribution within the patient.
- Incorporation into Treatment Planning: The accumulated dose is used to inform the optimization of the treatment plan for subsequent fractions, ensuring that the cumulative dose to the tumor meets the prescription while doses to OARs remain within tolerances [Heukelom, 2019].

Figure 2.8 shows a sequence of weekly cascade registrations in an in silico case [Heukelom, 2019], illustrating the use of deformable image registration and dose accumulation across therapy sessions. The parotid glands, which experience a nearly 25% reduction in volume, are tracked through this iterative "DVF chain" from the initial simulation to the final therapy session. In contrast, "dose back-projection" (green arrows) represents a reverse mapping, projecting the final therapy dose distribution back to the initial anatomical reference for comparative purposes.

The impact of these methods on dose distributions is further examined in Figures 2.9 B and C, where Figure B illustrates dose accumulation using DVFs across sessions, and Figure C demonstrates deformation back-projection. This comparison highlights the potential for parotid dose reduction if weekly dose adaptations were incorporated.

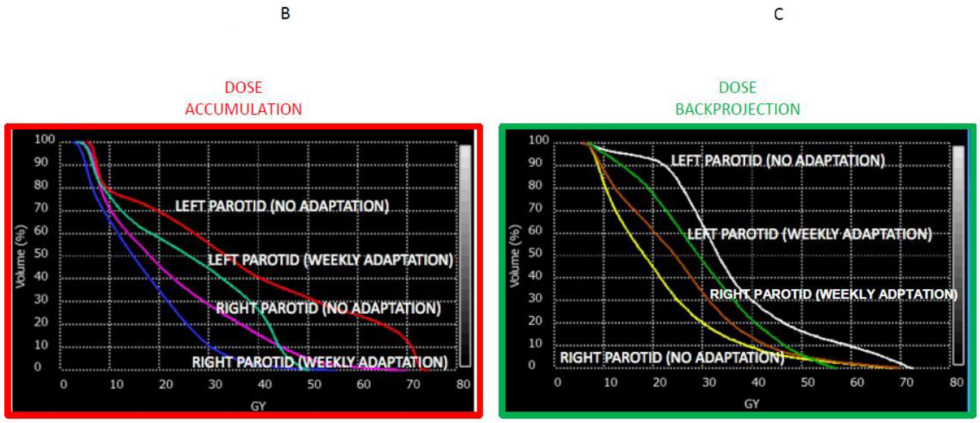

Figure 2.9: Comparative dose-volume histograms (DVHs) illustrating the effects of dose accumulation (B) and deformation backprojection (C) on parotid dose assessment. This figure highlights the impact of weekly dose adaptation [Heukelom, 2019] in minimizing irradiation to OARs

By continuously updating both the anatomy and the cumulative dose, cascaded adaptation aims to optimize the treatment plan dynamically over the entire course of therapy. This method ensures precise dose delivery to the target while minimizing exposure to OARs, taking into account both the immediate anatomical situation and the historical dose accumulation.

In an ideal scenario with unlimited computational resources, this approach would be preferred. Although theoretically feasible on several vendor systems, this data-intensive method has not yet been actively implemented. It represents a clear example of a "post-modern ART" application that requires additional support from vendors and manufacturers to become a reality [Heukelom, 2019].
Further technological and workflow improvements are essential for the wider adoption and effectiveness of ART in clinical settings. Achieving seamless integration of ART requires advancements in automation and innovative techniques that reduce manual intervention and streamline processes. Current developments and research efforts focus on these areas, aiming to make ART more efficient and accessible.

2.2.4 Advancements in Adaptive Radiotherapy Implementation

Automation in Segmentation and Treatment Planning

Automation plays a crucial role in making ART feasible in clinical practice by reducing the workload on clinicians and speeding up the adaptation process [Sonke, 2019].

Automated Segmentation Automated segmentation plays a key role in ART by enabling precise and efficient contouring of targets and OARs. Using atlas-based segmentation, a database of pre-delineated images is registered to the patient's images to generate automatic contours, which works well for OARs but may be less consistent for tumor targets due to anatomical variability [Teguh, 2011; Sharp, 2014]. Deep learning techniques, particularly convolutional neural networks (CNNs), offer fully automated segmentation with high accuracy and consistency, significantly reducing contouring time—an essential factor in online ART workflows [Zhu, 2019].

Automated Treatment Planning In treatment planning, automation is essential for generating high-quality plans efficiently, enabling rapid adaptation within session constraints. Knowledge-based planning systems leverage historical patient data to optimize plans for new cases, ensuring consistency and quality across treatments [Hussein, 2018]. Automated plan optimization, supported by advanced algorithms, adjusts treatment parameters to meet clinical objectives, reducing the need for manual intervention [Fan, 2019]. Additionally, automatic dose prediction tools use machine learning models trained on previous treatment plans to predict optimal dose distributions for individual patients, enhancing accuracy and speeding up plan generation [Nguyen, 2019].

Deformable Image Registration Deformable Image Registration (DIR) is a computational technique that aligns images from different times or modalities by accounting for anatomical deformations. In Adaptive Radiotherapy (ART), DIR plays a crucial role in two areas. First, it enables contour propagation by transferring delineated structures from planning images to those acquired during treatment, significantly reducing the need for manual re-contouring [Paganelli, 2018]. Second, DIR facilitates dose accumulation by mapping previously delivered doses onto the current anatomy, allowing for accurate cumulative dose assessment [Chetty, 2019].

Despite the uncertainties DIR can introduce, recent advancements have enhanced both its accuracy and reliability, supporting more precise ART applications.

Quality Assurance and Workflow Integration Implementing ART also depends on robust QA protocols and seamless workflow integration. Automated QA tools ensure treatment plan verification, confirming that adapted plans meet clinical standards before delivery [Hussein, 2018]. Workflow efficiency is improved through streamlined processes and software integration, reducing treatment times and resource demands, which is essential in high-volume clinical environments [Olberg, 2018]. Additionally, decision support systems, often AI-powered, aid in identifying patients who would benefit most from ART and help trigger adaptations based on set criteria [Hussein, 2018].

These advancements collectively enable the safe and effective implementation of ART, ultimately enhancing patient care.

2.2.5 High-Quality Fractional Imaging in ART

Adaptive Radiotherapy also fundamentally depends on high-quality imaging at each treatment fraction to accurately estimate the patient's 3D anatomy. Imaging quality directly influences critical aspects of ART, including image registration accuracy, dosimetry calculations, and clinical decision-making.

Precise alignment of images across different time points is essential for assessing anatomical changes and guiding adaptations to ensure accurate dose delivery. Furthermore, accurate representation of tissue densities and geometries is crucial for reliable dose calculations and treatment planning, while high-quality images enable clinicians to make informed decisions about the need for plan adaptations.

Challenges with Current Imaging Modalities

Two primary imaging modalities, CBCT and MRI, support ART. Each has distinct strengths and limitations that impact its suitability for ART applications.

CBCT CBCT integrated with linear accelerators provides in-room volumetric imaging, allowing clinicians to view the patient's anatomy immediately before treatment. However, CBCT images generally have lower quality than diagnostic CT, due to factors such as increased scatter radiation, which degrades image contrast, lower detector sensitivity, and the risk of motion artifacts due to slower acquisition times. CBCT's effectiveness for ART varies based on the disease site and the required image quality. It performs well in high-contrast areas, especially around bones, but its limitations become more evident in soft-tissue regions, such as the pelvis or the head and neck [Bissonnette, 2008]. These factors impair soft-tissue visibility and can compromise image registration and dose calculation accuracy. Indeed, CBCT often lacks accurate HU calibration, which is necessary for reliable dose calculations, leading many clinical protocols to rely on HU values from planning CT by deformable registration for dose accumulation, contouring and replanning [Thing, 2016].

MRI In-room MRI offers high soft-tissue contrast without the added ionizing radiation, making it valuable for ART. Its superior tissue contrast enhances tumor and OAR delineation, while real-time imaging enables dynamic monitoring of anatomical changes during treatment [Kupelian, 2014]. However, MRI lacks direct electron density information, required for dose calculations, which necessitates techniques like bulk density assignment or synthetic CT generation to approximate electron densities [Edmund, 2017]. Additionally, MRI often requires longer acquisition times, which can extend treatment sessions if not optimized.

Biplanar X-Rays Biplanar X-rays provide fast, low-dose, and cost-effective imaging for precise patient positioning and real-time motion correction but are limited in guiding adaptive radiotherapy. These systems capture two 2D images from orthogonal angles, offering accurate positional data but lacking the 3D detail needed for anatomical assessment and dose recalculation. Their limited field of view further restricts comprehensive 3D assessment, making biplanar X-rays more suitable for alignment during treatment sessions than for adaptive adjustments.

Implications for Adaptive Radiotherapy

Triggered Adaptation High-quality images are crucial in triggered adaptation to detect significant anatomical changes and guide plan adjustments accurately [Sonke, 2019]. Clear imaging reduces interobserver variability and enables consistent evaluations, while frequent repeat imaging captures relevant changes promptly.

Online Replanning Online replanning requires the highest imaging quality to enable precise contouring of target volumes and OARs for real-time adjustments. Advanced imaging allows accurate adaptation within each session, though current workflows often extend treatment times, highlighting the need for automation to streamline the process [Bohoudi, 2017; Hussein, 2018].

Geometric Uncertainties and Registration Accuracy Adaptive radiotherapy introduces additional geometric uncertainties related to imaging systems and image processing, which must be accounted for in the PTV margins to maintain treatment accuracy [Sonke, 2019]. In-room imaging systems, such as CBCT and MRI, need precise calibration to align with the linear accelerator's isocenter, yet calibration has finite precision, leading to uncertainties in anatomical positioning during treatment [Bissonnette, 2012]. Furthermore, in-room images may exhibit distortions due to hardware imperfections or magnetic field inhomogeneities in MRI, complicating precise anatomical localization [Weygand, 2016].

Rigid and deformable image registrations, used to align images across different time points, also introduce registration errors that can affect setup error assessments, contour propagation, and dose accumulation accuracy [Paganelli, 2018]. These compounded uncertainties may limit the effectiveness of ART by introducing dose delivery errors, underscoring the need for high imaging quality, precise registration, and consideration of these uncertainties in PTV margins to ensure accurate treatment coverage.

Enhancing Imaging Quality for ART

To better support ART, improvements in imaging quality are essential and can be achieved through both hardware and software strategies [Sonke, 2019].

Hardware enhancements include advanced detectors with higher sensitivity and resolution that can improve CBCT image quality, along with scatter reduction techniques such as anti-scatter grids that help reduce scatter radiation, thereby enhancing image contrast [Jin, 2010]. Optimized MRI sequences designed specifically for radiotherapy applications can also improve image quality and reduce acquisition time, which is critical in ART settings [Liu, 2015a].

Software innovations offer additional pathways to improve imaging quality. Advanced image reconstruction algorithms, like iterative reconstruction and compressed sensing, reduce noise and artifacts in CBCT images, while scatter correction algorithms mitigate the effects of scatter, further enhancing image clarity [Tian, 2011; Hansen, 2018]. Synthetic CT generation techniques, which create CT-equivalent images from MRI data, address the electron density information needed for dose calculations, making MRI more applicable in ART workflows [Edmund, 2017].

Deep learning plays a crucial role in enhancing image quality in ART, particularly through algorithms that generate synthetic CT (sCT) from CBCT data. Indeed, a growing body of research focuses on creating sCT from CBCT to improve patient positioning, delineation of OARs and dosimetric accuracy for CBCT-guided replanning. Algorithms like CycleGAN [Zhu, 2017] and U-Net [Ronneberger, 2015] have shown strong potential in transforming CBCT images into high-quality sCT, enabling precise dose calculations. Studies by Liang et al. and Kurz et al. highlight CycleGAN's effectiveness in generating sCT from CBCT, which is essential for accurate dose assessment in ART [Liang, 2019; Kurz, 2019]. Comparisons with U-Net and other algorithms have also shown favorable results for CBCT correction and sCT generation across different anatomical sites and treatment modalities [Landry, 2019; Thummerer, 2020].

Following the ALARA (As Low As Reasonably Achievable) principle, established by the International Commission on Radiologic Protection (ICRP) in 1977 [Radiological Protection, 1977; Hendee, 1986], which states that radiation exposure should be minimized to the lowest possible levels while still achieving the required diagnostic or therapeutic quality, recent efforts have focused on reducing fractional imaging doses while preserving image quality. Techniques include the use of CNNs, such as U-Net, to enhance low-dose CBCT images for head and neck radiotherapy [Yuan, 2020], and GANs, like CycleGAN [Zhu, 2017], for generating synthetic CTs from low-dose CBCT in adaptive radiotherapy [Gao, 2021]. CycleGAN and CUT [Park, 2020], have also been employed to correct artifacts in 4D CBCT images [Dong, 2022]. Recently, Chan et al. evaluated the minimum CBCT dose needed for pelvic synthetic CT generation using CycleGAN and CUT algorithms, concluding that a 25% dose is necessary for accurate VMAT dose calculations and reliable organ delineation in ART [Chan, 2024]. Figure 2.10 illustrates this pipeline.

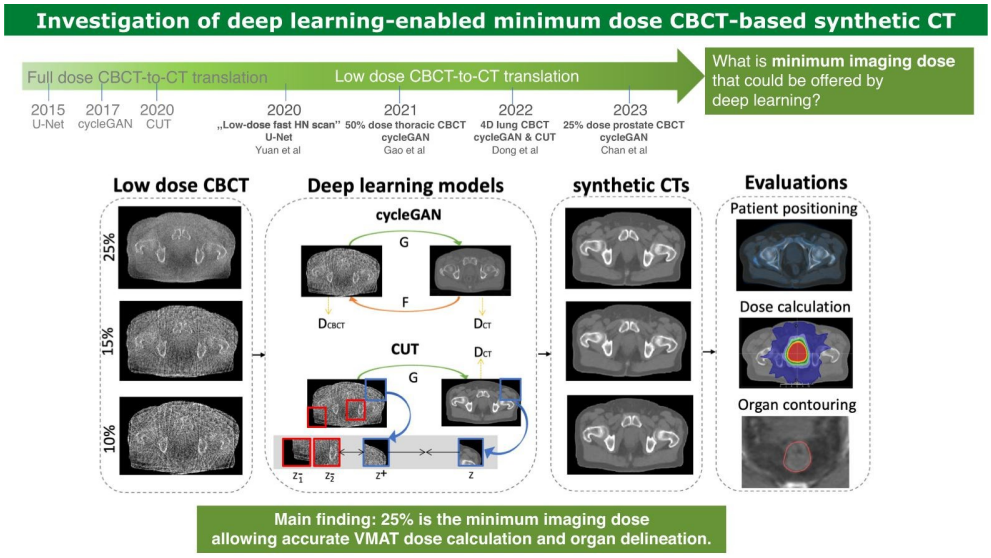

Figure 2.10: Pipeline for generating synthetic CT from low-dose CBCT using CycleGAN and CUT architectures, evaluated at dose reductions of 25%, 15%, and 10%, for accuracy in patient positioning, dose calculation, and organ contouring [Chan, 2024].

Researchers, including [Chan, 2024], have explored the shift from full-dose to low-dose CBCT-to-CT translation, as illustrated in Figure 2.10, raising a key question: what is the minimal dose that deep learning can achieve to guide Adaptive Radiotherapy?

Could we push to extreme ? Down to 1% ? To 2 X-rays? This thesis seeks to redefine imaging in Adaptive Radiotherapy by envisioning a system that reconstructs a 3D image from only two biplanar X-rays. Such a breakthrough would enable minimal radiation exposure, instant acquisition, and minimum cost, offering the potential for CT-quality imaging with unparalleled speed and precision (2.11).

However, 3D reconstruction from biplanar X-rays is a highly ill-posed inverse problem due to the extreme sparsity of projections, leading to significant ambiguity and a wide

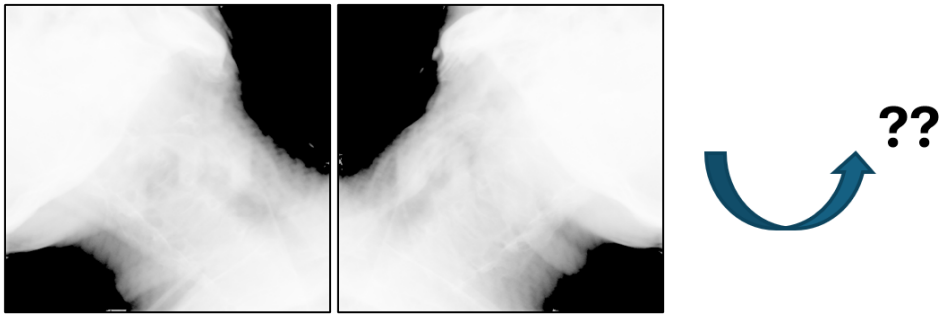

Figure 2.11: Biplanar X-rays of the head and neck from ExacTrac: Can we reconstruct 3D images from just two projections to enable adaptive radiotherapy with minimal dose, low cost, and fast acquisition ?

range of possible reconstructions. Furthermore, the narrow FOV of biplanar X-rays can restrict comprehensive anatomical assessment and limit dosimetric accuracy needed for adaptive radiotherapy.

These limitations emphasize two main challenges: achieving accurate 3D reconstruction from just two projections and expanding the FOV to include critical surrounding structures. This thesis primarily addresses the highly ill-posed challenge of 3D reconstruction from biplanar X-rays. Finally, we also present how we could reconstruct areas outside the initial field of view.

2.3 Conclusion

This chapter reviewed advancements in Image-Guided Adaptive Radiotherapy for head and neck cancer, emphasizing its role in enhancing treatment precision. By enabling real-time adjustments, IGART helps to adapt to significant anatomical changes during therapy, balancing therapeutic efficacy with quality of life. From offline to complex online adaptive strategies, these techniques improve dose delivery accuracy while minimizing exposure to healthy tissues. Innovations in imaging quality, automation, and deformable registration continue to advance IGART, paving the way for low-dose imaging and rapid 3D reconstruction.

Currently, no reliable methods exist to create the robust 3D images required for adaptive radiotherapy solely from biplanar X-rays. In the following chapters, we examine the feasibility of this vision, investigating whether adaptive radiotherapy could be guided by just two projections—a low-dose, instant, and low-cost solution for best outcomes.

Chapter 3

3D Reconstruction from Biplanar X-Rays

Reconstructing three-dimensional (3D) anatomy from biplanar X-rays offers a promising, low-dose, and cost-effective solution for guiding adaptive radiotherapy. However, with only two projections, the reconstruction problem becomes highly ill-posed, making it challenging to accurately capture complex anatomical structures.

This chapter addresses this challenge by first exploring the theoretical foundations of tomography and the nature of ill-posed inverse problems. We discuss how techniques like compressed sensing, combined with generative models, leverage learned anatomical priors to effectively constrain the solution space, enabling accurate reconstructions from limited measurements.

We introduce **X2Vision**, a generative model-based approach for reconstructing 3D head and neck anatomy from biplanar X-rays. By integrating anatomical priors through a generative model trained on 3D CT scans, X2Vision captures a low-dimensional manifold of plausible anatomies. Optimizing within this latent space allows us to produce a 3D volume that aligns with both the anatomical priors and the input projections, achieving greater accuracy and robustness than traditional methods.

This chapter details the core strategies behind X2Vision and presents experimental results demonstrating its potential for adaptive radiotherapy applications, including precise patient positioning and dosimetry simulations. We conclude with a discussion of current limitations and future directions for improving this approach.

Contents

3.1 Theoretical Foundations of Tomography

3.1.1 Tomography

Tomography, from the Greek words tomos (meaning "cut" or "section") and graphia ("writing" or "representation"), refers to the imaging process used to reconstruct an object's internal structure by analyzing external measurement data [Kak, 2001]. This method produces cross-sectional images that represent slices of an object, revealing its internal characteristics. Tomography has diverse applications, from non-destructive testing (examining the interior of materials without damage) to fields such as geophysics, astrophysics, and especially medical imaging, where it is essential for diagnosing and treating diseases by visualizing internal anatomy.

The basic principle of tomography is to measure radiation that interacts with the object by emission, transmission, or reflection. These measurements indirectly reveal internal properties, like tissue density or material composition, which are then reconstructed into an interpretable image.

Medical Tomography

In medicine, tomography is central for producing detailed images of the human body, and different imaging devices are used based on specific diagnostic needs:

- CT (Computed Tomography): Uses X-rays to assess tissue density by measuring radiation absorption.
- SPECT (Single Photon Emission Computed Tomography): Captures gamma rays emitted by radiotracers injected into the body.
- PET (Positron Emission Tomography): Employs positron-emitting radiotracers to visualize metabolic activity.
- Optical Tomography: Uses reflected light to image structures.

At the heart of medical tomography is the acquisition of projections—radiation measurements taken from various angles around the body. These projections allow for a 3D reconstruction of internal structures. Medical tomography divides into two main types: transmission tomography (measuring radiation passing through the body) and emission tomography (measuring radiation emitted from within the body).

Transmission Tomography Transmission tomography employs an external radiation source, like X-rays, to pass through the body, with tissue density and composition attenuating the radiation intensity detected on the other side. This attenuation pattern provides critical information about internal structures and can be mathematically described by the Beer-Lambert law [Beer, 1852], which models how radiation diminishes as it passes through a medium:

$$I = I_0 \exp(-\mu L), \tag{3.1}$$

where I_0 is the initial intensity of the radiation, I is the detected intensity after passing through the tissue, μ is the attenuation coefficient (typically expressed in cm^{-1}), L is the thickness of the material. For example, in water, at 140 keV, the attenuation coefficient μ is 0.15 cm^{-1}. After traversing 20 cm of water, only 5% of the original photon intensity remains [Buvat, 2006].

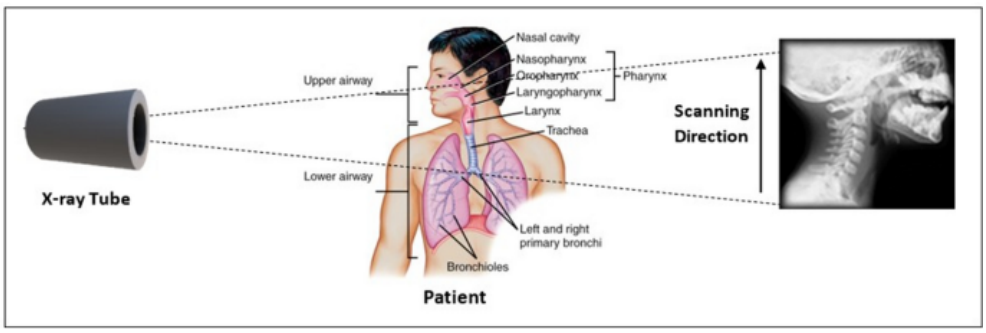

Figure 3.1: Emission and transmission of X-rays through the head and neck region to produce 2D projection image. The X-ray tube generates beams that penetrate the patient, and the detector capture the attenuated rays after they pass through various tissues. The collected data forms a 2D projection that reflects the internal anatomical structures. Adapted from [Maken, 2023].

Emission Tomography Emission tomography, including techniques like PET and SPECT, detects radiation emitted from within the body, typically from radiotracers administered to the patient. In PET, positrons from the tracer interact with electrons, producing gamma rays that detectors capture. In SPECT, gamma cameras detect photons emitted directly from the radioactive decay of the tracer. The main challenge in emission tomography is reconstructing the spatial distribution of the radiotracer within the body based on this detected radiation.

X-Ray Imaging

X-rays are high-energy photons, a form of electromagnetic radiation with short wavelengths that enable them to penetrate various materials, including human tissue. As a core component of medical imaging, X-rays are used in techniques such as radiography and CT, playing a critical role in reconstructing internal structures from external measurements.

X-rays are typically generated by accelerating electrons and colliding them with a metal target (commonly tungsten), which results in the emission of photons. Figure 3.2 illustrates this. These photons are directed at an object, such as the human body, and their interaction with the tissues is the foundation of X-ray imaging.

The key interactions between X-rays and matter include:

- Absorption (Photoelectric Effect): X-ray photons are absorbed by atoms in the tissue, with their energy being used to eject electrons from the atom. This process is heavily influenced by the material's atomic number, which is why denser materials like bone absorb more X-rays.
- Rayleigh Scattering: This occurs when X-ray photons are scattered in a random direction without a change in their wavelength or energy, primarily by atoms with low atomic numbers. The scattering is stochastic, meaning the direction of the scattered photons is unpredictable, though their energy remains unchanged.

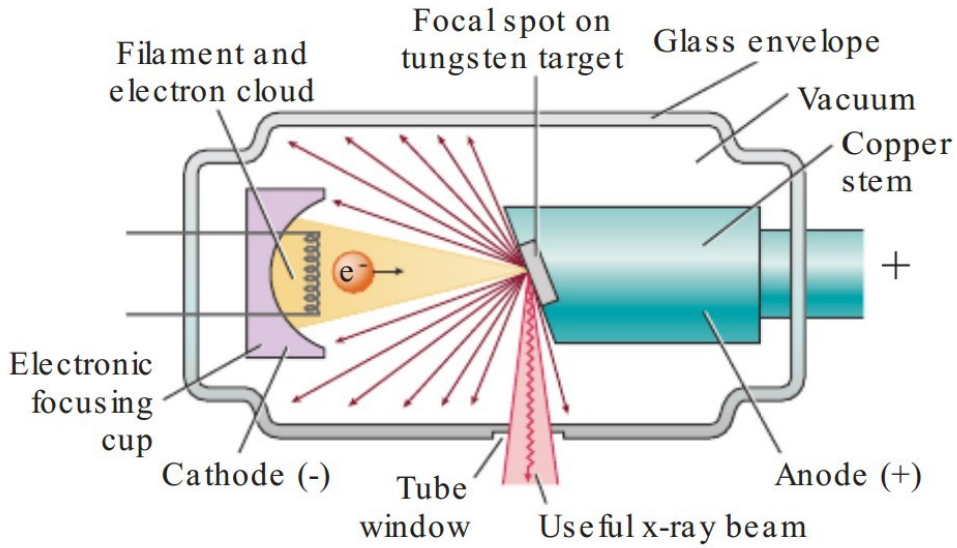

Figure 3.2: X-ray generation. Electrons are accelerated and collided with a metal target, resulting in the emission of X-ray photons, which are directed at the object for imaging. From [Behin Negareh Co, 2023].

- Compton Scattering: In this process, X-ray photons collide with loosely bound electrons, causing the photon to lose some of its energy and change direction randomly. Compton scattering is more prevalent in soft tissues and often leads to image degradation by introducing scattered photons that blur the final image.

X-rays that pass through the body and reach the detector without interacting with tissues contribute to image formation, while those that are absorbed or scattered do not. As X-rays travel through different tissues—such as bone, muscle, or air—they experience varying degrees of attenuation (reduction in intensity), which creates contrast in the resulting images. Typically, X-ray images display attenuation as an inverted representation of transmission, where denser structures, like bone, appear white, indicating higher attenuation. Figure 3.1 illustrates the emission of X-rays through the human head and neck area to create 2D projection of the anatomical structures.

Low-energy X-rays, typically used in diagnostic imaging, are designed to minimize radiation exposure while providing sufficient detail. CT imaging, however, requires multiple projections from various angles to generate detailed 3D images, which cumulatively increase the overall radiation dose. In contrast, the higher-energy X-rays produced by linear accelerators (Linacs) in radiotherapy deliver a much higher dose of radiation, as they are intended to treat, rather than image, the target tissue.

The Beer-Lambert Law describes the relationship between the intensity of an X-ray beam before and after passing through the body. It models how X-rays are attenuated as they interact with different materials in the body. For cases involving multiple materials

such as in human body (tissues, bones, air) and different energy levels, the equation becomes more complex than 3.1. It accounts for the sum of attenuation across different materials and energy states. The resulting attenuated intensity can be expressed as [Unberath, 2018; Peng, 2021; Bushberg, 2011]:

$$I = \sum_{E} I_0 e^{-\sum_m \mu(m,E)t_m} + S_E + \text{noise}, \tag{3.2}$$

where $\mu(m, E)$ is the linear attenuation coefficient for material m at energy level E, which is known from standard measurements [Hubbell, 1995], t_m is the thickness of material m, I_0 is the initial X-ray intensity, S_E represents the scatter estimation term, accounting for scattered X-rays, and the final term represents noise in the measurement.

As a result, X-rays involve several factors, including:

- Ambiguity: X-ray measurements are the result of attenuation integration across the body, which makes them very ambiguous as X-rays travel through the body. They compress 3D information into a flat 2D image. This makes it very hard to determine the exact location and span of structures inside the body.
- Noise: Noise in X-ray imaging comes from the stochastic nature of photon emission and detection, typically modeled as Poisson noise. This noise becomes more pronounced in low-dose X-ray or CT imaging, where fewer photons are used.
- Scattering: Scattering is a major source of image degradation, as scattered photons lead to blurred images and reduced contrast. The accurate modeling of scattering effects is important for improving image reconstruction algorithms.
- Multi-Energy Spectra: X-ray beams are generally polychromatic, comprising a range of energies. Different energy levels interact uniquely with various tissues, leading to complex absorption patterns that complicate the modeling of X-ray interactions. For accurate image reconstruction, spectrum-based ray tracing techniques are essential. Figure 3.3 illustrates a typical X-ray energy spectrum, highlighting both Bremsstrahlung and characteristic X-ray peaks.

Tomographic Reconstruction

Tomographic reconstruction aims to reconstruct internal 3D structures from multiple projections, formulated as an inverse problem. In this context, an inverse problem involves determining an object's unknown internal structure based on external measurements. Specifically, in transmission tomography, the objective is to determine the 3D spatial distribution of the attenuation coefficient, $\mu(l)$, from the measured projections [Kak, 2001]:

$$\ln\left(\frac{I_0}{I}\right) = \int_0^L \mu(l)\, dl, \tag{3.3}$$

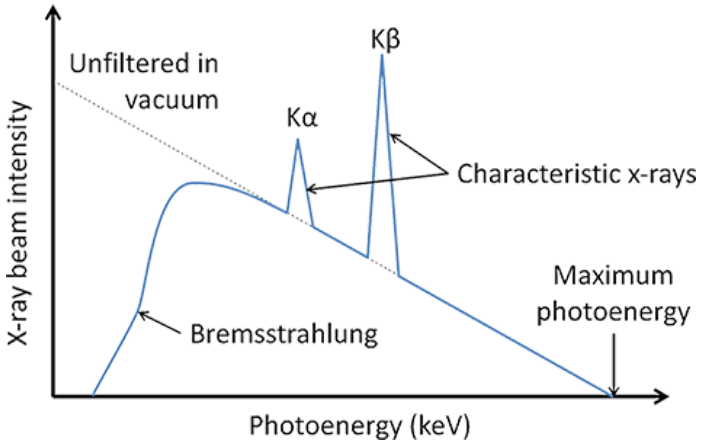

Figure 3.3: Typical X-ray spectrum showing the contributions of Bremsstrahlung radiation and characteristic X-ray peaks ($K\alpha$ and $K\beta$). The unfiltered curve represents the full spectrum in vacuum, with maximum photoenergy corresponding to the peak X-ray intensity. Adapted from [Radiology Cafe, 2023].

where I_0 is the initial radiation intensity, I is the detected intensity after passing through the tissue, $\mu(l)$ represents the attenuation coefficient at position l along the radiation beam path, and L is the length of the beam path through the tissue.

To obtain a 3D representation of an object, multiple 2D projections must be collected from various angles. The process of reconstructing a 3D volume from these projections involves solving an inverse problem where the goal is to estimate the 3D structure that most accurately explains the measured projections. We will explore methods for reconstructing 2D slices from 1D projections. By combining these 2D slices, we can create 3D reconstructions.

Figure 3.4 illustrates the process of 3D tomographic reconstruction from multiple cone-beam X-rays.

Ill-Posed Problem Tomographic reconstruction is ill-posed. An ill-posed problem in mathematics refers to a problem that does not satisfy one or more of the conditions necessary for a well-posed problem, as defined by Hadamard in 1923: (i) the solution must exist, (ii) the solution must be unique, and (iii) the solution must depend continuously on the input data. Tomographic reconstruction is inherently ill-posed due to several factors[Herman, 2009; Buvat, 2006]:

- Non-uniqueness: A finite set of projections can correspond to multiple possible internal structures. This means that different 3D structures can produce the same projection data, leading to ambiguity in the reconstructed image.
- Instability: Even small changes or errors in the projection data can cause large differences in the reconstructed image. This instability makes the reconstruction process highly sensitive to measurement inaccuracies and noise.

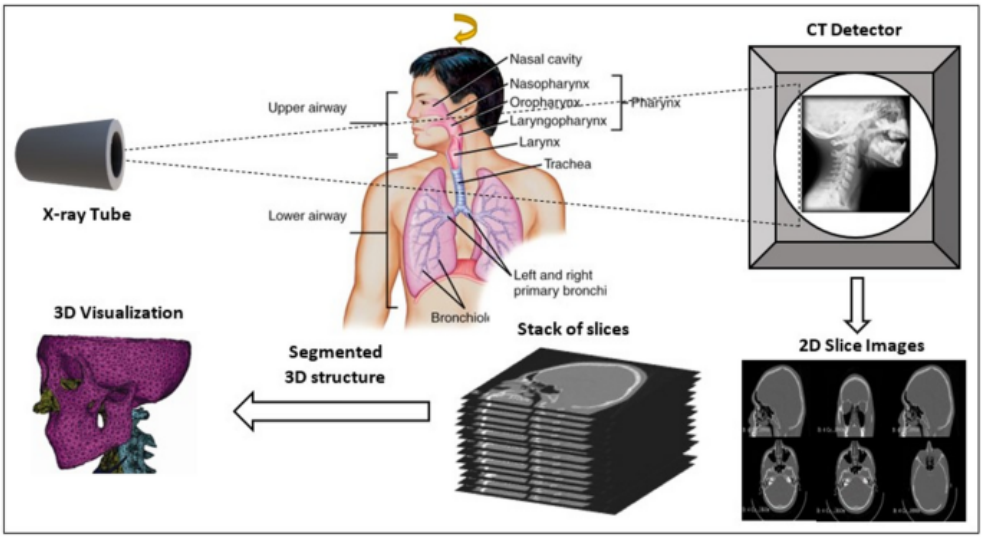

Figure 3.4: Process of obtaining a 3D representation from X-ray imaging. Multiple 2D projections are captured from different angles using an X-ray tube and CT detector setup. These projections are used to reconstruct a stack of 2D slice images, which are then combined to form a 3D structure. The final 3D visualization enables detailed examination of anatomical regions. From [Maken, 2023].

Radon Transform Tomographic reconstruction is based on principles established by Johann Radon in 1917 through the Radon Transform [Radon, 2005], which relates projections of a two-dimensional object to its original structure. The goal is to estimate the internal configuration of an object by integrating its projections taken from various angles.

The Radon Transform of a two-dimensional function $f(x, y)$ is defined as the integral of f along a straight line L:

$$Rf(p, \theta) = \int_L f(x, y)\, dl. \tag{3.4}$$

Here, the line $L = L(\theta, p)$ is defined by:

$$p = x \cos\theta + y \sin\theta, \quad \theta \in [0, 2\pi), \tag{3.5}$$

where p is the perpendicular distance from the origin to the line L and θ is the angle between L and the x-axis.

For a fixed angle θ, as p varies over all real numbers, $Rf(p, \theta)$ constitutes a projection of $f(x, y)$. Collecting these projections for all angles $\theta \in [0, 2\pi)$ results in a dataset known as a *sinogram*. A sinogram visually represents the projection data acquired from multiple angles for a specific slice of an object, with each line corresponding to a projection at a particular angle.

The term "sinogram" arises from the sinusoidal paths traced by the projections of point objects as θ varies. By gathering projections at multiple angles, the original function $f(x, y)$ can be reconstructed using inverse Radon Transform techniques [Kak, 2001], which

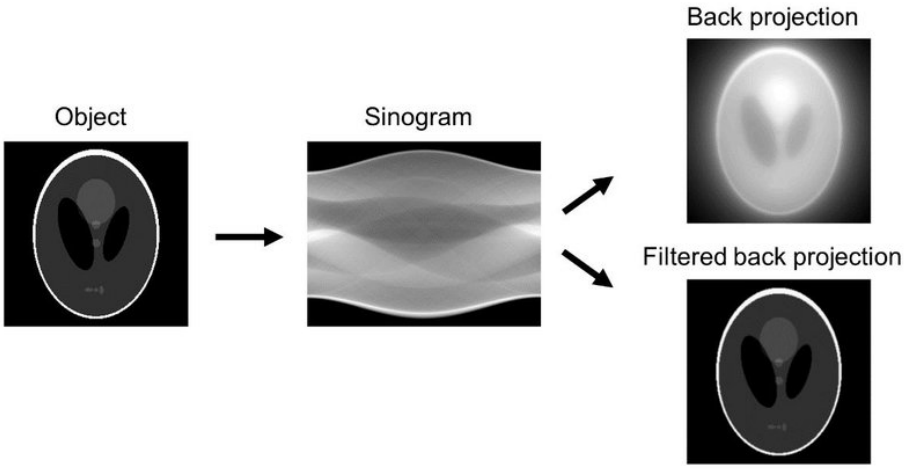

Figure 3.5: Illustration of Filtered BackProjection of Shepp-Logan phantom: sinogram representing projection data over all angles, backprojection result showing blurring, and filtered backprojection result, which accurately reconstructs the original object. From [Vamvakeros, 2017].

is fundamental in tomographic imaging methods such as computed tomography (CT).

3.1.2 Approaches to Tomographic Reconstruction

Reconstruction techniques can be divided into two main categories [Herman, 2009]:

- Analytical Methods: These rely on direct inversion techniques.
- Iterative Methods: These involve refining an initial guess of the image through an iterative process, minimizing errors between measured and generated projections.

Analytical Methods

Filtered BackProjection Filtered BackProjection (FBP) [Herman, 2009; Kak, 2001] is one of the most widely used analytical techniques for image reconstruction. The core idea is to first apply a filtering process to the projections and then backproject them across the image space to reconstruct the original object.

Backprojection alone simply spreads the projection data across the image but results in blurring and artifacts. To counteract this, FBP introduces an initial filtering step in the frequency domain, which reduces blurring and improves reconstruction accuracy. Mathematically, backprojection can be represented as:

$$f^*(x,y) = \int_0^{\pi} p(u,\theta)\, d\theta, \tag{3.6}$$

where $f^*(x,y)$ is the reconstructed approximation. However, backprojection alone does not yield an exact inverse of the Radon Transform. To achieve accurate reconstruction, filtered backprojection applies a filter derived from the central slice theorem, which

links projections to the object's representation in Fourier space, enabling precise image reconstruction.

Figure 3.5 illustrates this process using the Shepp-Logan phantom [Shepp, 1974], a standard test image in tomography that simulates the cross-sectional structure of a human head with different materials and shapes. The figure shows the acquired sinogram, the result of simple backprojection with visible blurring, and the outcome of filtered backprojection, which provides a much closer approximation to the original object when the sinogram includes projections from all angles.

The Central Slice Theorem, also known as the Fourier Slice Theorem, is crucial to the FBP process [Kak, 2001]. It states that the 1D Fourier transform of a projection at a given angle corresponds to a slice through the 2D Fourier transform of the original object at the same angle. This relationship allows the reconstruction process to be simplified by performing filtering in the frequency domain.

Filters are crucial in FBP for balancing image resolution and noise [Buvat, 2006; Herman, 2009]. Common filters include the Ramp Filter, which enhances high-frequency details but may increase noise, the Hann Filter, which smooths noise while sacrificing some high-frequency information, the Gaussian Filter, which provides strong noise reduction at the cost of blurring fine details, and the Butterworth Filter, which allows adjustable control over the trade-off between resolution and noise. The choice of filter depends on the desired balance of clarity and noise reduction in the image.

Typical CT geometries include parallel-beam, fan-beam, and cone-beam configurations. Fan-beam CT captures fewer slices per projection, offering higher precision but requiring more projections to cover the entire volume. In contrast, cone-beam CT captures multiple slices per projection, reducing the total number of projections needed but increasing noise and artifacts due to scatter. This scatter can become significant in cone-beam CT, sometimes exceeding primary radiation, which leads to reduced contrast and introduces artifacts like streaking and cupping [Graham, 2007].

The Figures 3.6 and 3.7 highlight differences in CT geometries and the effects of fan-beam and cone-beam geometries in the quality of FBP.

One of the most widely applied adaptations of FBP for cone-beam CT is the Feldkamp-Davis-Kress (FDK) method [Feldkamp, 1984a]. The FDK algorithm extends FBP to handle the cone-beam geometry by incorporating a weighted backprojection and a ramp filter. In this approach, each projection is corrected for the cone angle and weighted accordingly, which allows accurate reconstruction of 3D objects from cone-beam CT data.

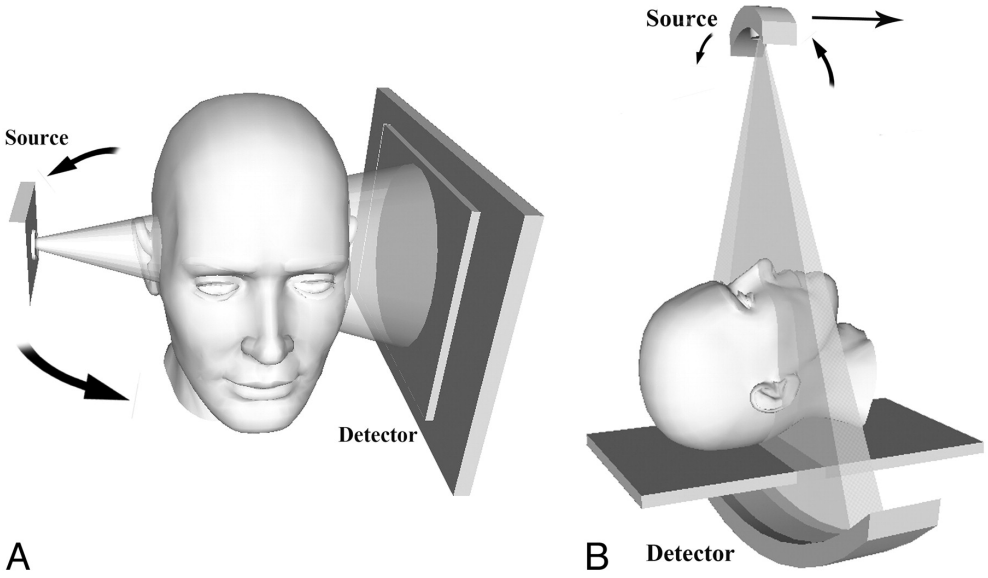

Figure 3.6: Comparison of fan-beam and cone-beam geometries. Fan-beam CT provides higher spatial accuracy by capturing fewer slices per projection, while cone-beam CT covers more volume per projection but is prone to increased noise and artifacts due to scatter. Adapted from [Miracle, 2009]

Discussion Analytical methods offer several advantages. They are known for their speed and simplicity, being computationally efficient and easy to implement, which makes them the preferred choice in many clinical and industrial applications. Additionally, these methods are linear, meaning that if the projection values are doubled, the reconstructed values will also double, preserving proportionality and making the results predictable [Herman, 2009; Buvat, 2006].

However, there are notable limitations. Analytical methods often rely on assumptions of perfect resolution and noise-free data, which are rarely true in practice. FBP is particularly sensitive to noise, and even minor errors in the projection data can introduce artifacts like streaking. Furthermore, these methods tend to ignore complex physical phenomena such as attenuation, scattering, and detector imperfections, which can lead to inaccuracies in the final image. When the number of projections is limited, analytical methods struggle to produce accurate reconstructions, leading to poor image quality.

These shortcomings have led to more robust iterative methods, which offer greater flexibility by incorporating physical effects and noise into the reconstruction process. Iterative methods can also integrate regularization, making them better suited for sparse sampling and challenging conditions, so resulting in higher-quality images.

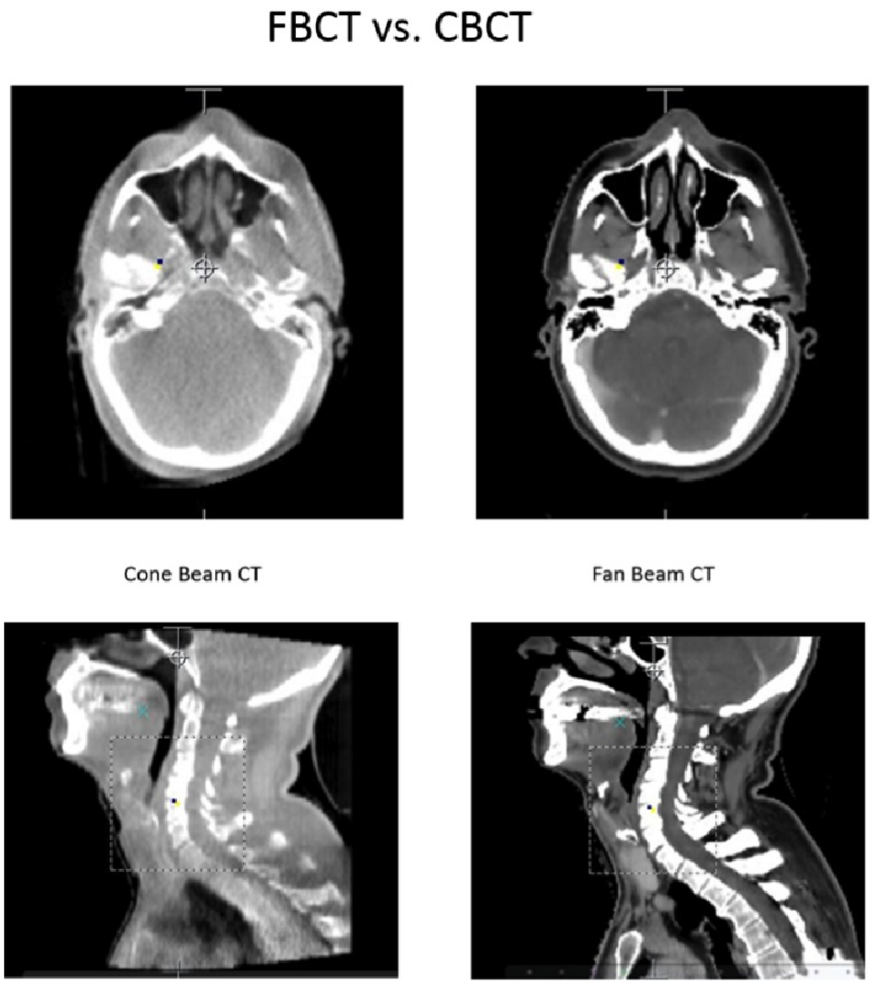

Figure 3.7: Illustration of reconstruction with fan-beam CT (FBCT/CT) and cone-beam CT (CBCT) geometries. CBCT, due to its cone-beam setup, generally has a lower signal-to-noise ratio and is more susceptible to scatter artifacts, reducing image contrast and detail compared to FBCT. Adapted from [Lechuga, 2016].

Iterative Reconstruction Methods

In contrast to analytical methods, iterative reconstruction methods [Gordon, 1970; Hansen, 2006]approach the tomographic problem by progressively refining an estimate of the object being imaged. The iterative process aims to minimize the difference between the measured projections and those generated from the current estimate of the object.

The core problem is represented as a system of linear equations in $\mathbb{R}^n$:

$$p = Rf, \tag{3.7}$$

where $p \in \mathbb{R}^m$ represents the acquired projections, $R \in \mathbb{R}^{m \times n}$ is the projection operator, and $f \in \mathbb{R}^n$ is the image to be reconstructed.

The goal is to solve f, given the known p and R.

The Projection Operator R The projection operator R describes the image formation process, linking the object to the measured projections. This operator encompasses two key aspects:

- Geometric Modeling: This refers to how each voxel contributes to the projection data, taking into account the shape and geometry of the imaging system. It includes factors such as the source-to-detector distance, the distance to the isocenter, acquisition angles, beam geometry (e.g., fan-beam or cone-beam), sensor size, and the induced field of view at the isocenter.
- Physical Modeling: This accounts for factors like attenuation, scattering, and detector response, ensuring accurate representation of how the system acquires the data.

There are two primary classes of iterative reconstruction methods: Algebraic and Statistical [Herman, 2009].

Algebraic Methods These methods iteratively solve a system of linear equations, aiming to minimize the error between the measured projections and the projections generated from the current image estimate. Commonly used algorithms include:

- ART (Algebraic Reconstruction Technique) [Gordon, 1970]: This method refines the image estimate by iteratively applying a correction factor based on individual projections. Each iteration corrects the estimate along a single projection direction to reduce the corresponding projection error.
- SIRT (Simultaneous Iterative Reconstruction Technique) [Baker, 1985]: SIRT updates the image estimate by considering the full set of projections simultaneously in each iteration, calculating an average correction across all directions. This approach often results in smoother convergence compared to ART, especially in cases of noisy or incomplete data.

These methods focus solely on minimizing the residual error between measured and estimated projections but do not inherently model noise or statistical variations in the data.

Statistical Methods Statistical methods account for the probabilistic nature of measured data, aiming to maximize the likelihood that the estimated image aligns with the observed data. These methods are particularly beneficial in noisy or low-dose imaging.

- **MLEM (Maximum Likelihood Expectation Maximization)** [Shepp, 1982]: Assumes the data follows a Poisson distribution. MLEM iteratively refines the image by maximizing the likelihood function, offering high-quality reconstructions but with slower convergence.
- **OSEM (Ordered Subset Expectation Maximization)** [Hudson, 1994]: An accelerated version of MLEM, OSEM divides the data into subsets to update the image faster, improving convergence speed at the cost of potential bias if not regularized.
- **RAMLA (Row Action Maximum Likelihood Algorithm)** [Browne, 1996]: Extends OSEM with a relaxation parameter to control noise, achieving smoother convergence and making it suitable for high-noise scenarios.

Regularization Regularization has been introduced to guide the solution toward a plausible outcome, often by incorporating prior information about the object imaged to reconstruct. Compared to analytical methods, regularization can be easily incorporated in iterative methods. Regularization plays a critical role in iterative reconstruction by preventing the solution from overfitting noisy or incomplete data. Regularization methods include [Buvat, 2006; Hansen, 2006]:

- A priori modeling: Incorporating prior knowledge, such as anatomical information from pre-captured MRI or CT, to guide the reconstruction.
- Empirical methods: Techniques like early stopping of iterations or applying post-filtering to reduce noise.
- Variational regularization: Minimizes both the projection error and a regularization term (such as sparsity, smoothness) that penalizes unlikely solutions, balancing data fidelity and regularization. From a Bayesian perspective, it is often framed as Maximum A Posteriori (MAP) estimation, where the regularization term represents a prior distribution.

Discussion Iterative methods present several significant advantages over analytical approaches [Thibault, 2007; Beister, 2012; Buvat, 2006]. A key benefit is their flexibility in modeling complex physical phenomena, such as attenuation, scattering, and motion, within the projection operator. This capability allows iterative methods to compensate for various effects that would otherwise degrade image quality, making them particularly useful in scenarios where these phenomena are significant. Moreover, statistical iterative methods are highly effective at handling noise, especially in low-count or low-irradiating imaging contexts. This one of the reasons iterative methods outperform analytical methods when dealing with incomplete or noisy datasets, where FBP may struggle to produce accurate results.

However, iterative methods come with challenges. They are computationally intensive, requiring multiple iterations to converge, though advances in computing power have mitigated this issue. Their non-linearity complicates image interpretation and optimization compared to the more straightforward FBP. Additionally, controlling the number of iterations is crucial—too few result in incomplete reconstructions, while too many amplify noise and introduce artifacts. While iterative methods are less prone to streak artifacts and better suited for noisy data, they can introduce new noise patterns if not properly regularized and remain slower than FBP. Research now favors iterative methods due to their flexibility in modeling complex physical and statistical processes. With increasing computational power, these techniques are becoming more practical for use in clinical and research settings.

Tomography with Very Few X-rays Tomographic reconstruction, even with hundreds of projections, is fundamentally an ill-posed problem, and current methods provide approximations rather than perfect reconstructions of internal structures. Despite this, these approximate solutions have proven sufficiently accurate for widespread use in medical practice.

However, with extremely sparse data—such as only two projections—traditional methods fail entirely, resulting in coarse reconstructions where internal boundaries cannot be clearly distinguished, as illustrated in Figure 3.8, which shows FBP reconstructions of the Shepp-Logan phantom with varying numbers of projections. In such cases of extreme data scarcity, much stronger regularization is needed to produce any useful reconstruction. Effective regularization in these scenarios often relies on incorporating anatomical constraints, such as learning typical anatomical structures, to guide and improve the accuracy of the reconstruction process.

This is where deep learning has transformed the field. By learning statistical distributions of realistic anatomies, deep learning models provide robust regularization, enabling much more accurate reconstructions from minimal data. In the following section, we will explore traditional approaches to addressing ill-posed inverse problems and examine how deep learning has enabled viable reconstructions from very few measurements.

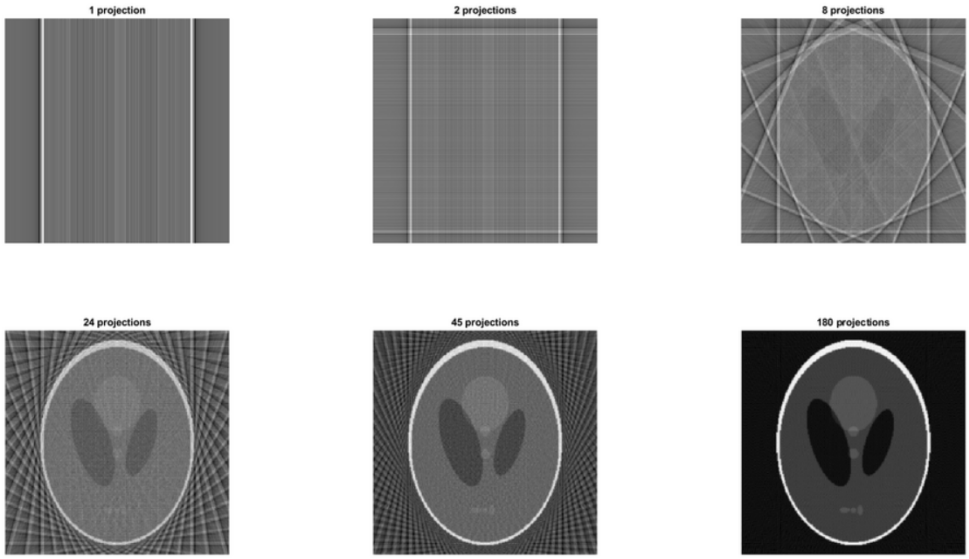

Figure 3.8: Filtered Back-Projection reconstructions of the Shepp-Logan phantom with varying numbers of projections. With only one or two projections, the reconstruction is highly ambiguous, lacking clear structure. Dozens of projections are required to provide enough constraints for recognizable structures to emerge. From [DotEagle, 2023].

3.2 Solving Ill-Posed Inverse Problems

3.2.1 Inverse Problem

Reconstructing 3D structures from only two X-ray projections poses a highly challenging and ill-posed inverse problem. Inverse problems involve estimating an unknown object or signal based on measurements that are typically indirect, incomplete, or noisy. This is generally modeled as:

$$y = Ax + \eta, \tag{3.8}$$

where $y \in \mathbb{R}^m$ represents the observed data (e.g., X-ray projections), $A \in \mathbb{R}^{m \times n}$ is the measurement matrix (e.g., the projection model), $x \in \mathbb{R}^n$ is the unknown structure to be reconstructed (e.g. the 3D structure), and $\eta \in \mathbb{R}^m$ accounts for measurement noise or uncertainties. The forward model A may include processes such as downsampling, motion blur, artifacts, masking, or projection [Kaipio, 2006; Bora, 2017].

Solving inverse problems is difficult because the forward process is typically non-invertible, meaning x cannot be directly recovered from y. This challenge arises in various imaging tasks, such as deblurring, inpainting, and superresolution, where information loss in the forward process renders the problem ill-posed. Consequently, there are often multiple possible solutions that match the observed data.

To address this, traditional reconstruction methods, like those discussed in the previous section, minimize a cost function that balances:

- A data fidelity term $\|y - Ax\|_2^2$, ensuring that the reconstructed structure x is consistent with the observations y.
- A regularization term, which incorporates prior knowledge about the underlying structure. This term encourages desirable properties in the solution, such as sparsity, smoothness, or anatomical structures in medical imaging. Regularization is crucial for constraining the solution space, thus reducing ambiguity in the reconstruction.

Another common way to represent the inverse problem is from a probabilistic perspective, using a Bayesian framework [Kaipio, 2006]. This approach enables us to incorporate prior knowledge about the structure of the solution, which is especially useful in ill-posed problems. Under this view, we aim to estimate the posterior distribution of the image x given the observations y, denoted as $p(x|y)$. According to Bayes' theorem, the posterior can be expressed as:

$$p(x|y) \propto p(y|x) \cdot p(x), \tag{3.9}$$

where $p(y|x)$ is the likelihood, representing how likely the observed data y is given the image x, and $p(x)$ is the prior, which encodes prior knowledge about plausible structures of x. This formulation allows us to regularize the reconstruction with domain-specific information.

One common approach is to find the Maximum A Posteriori (MAP) estimation, which seeks the image x that maximizes the posterior distribution:

$$x_{\text{MAP}} = \arg\max_x \, p(x|y). \tag{3.10}$$

Equivalently, the MAP estimate can be obtained by minimizing the negative log of the posterior:

$$x_{\text{MAP}} = \arg\min_x \, [-\log p(y|x) - \log p(x)]\,. \tag{3.11}$$

This formulation mirrors the traditional approach of minimizing a cost function with a data fidelity term (log-likelihood) and a regularization term (log-prior), allowing the prior to incorporate domain-specific knowledge, such as typical anatomical structures in medical imaging. This Bayesian framework provides a principled way to balance observed data with realistic, prior-based constraints.

3.2.2 Compressed Sensing

Challenges in solving ill-posed inverse problems often arise because they lead to an underdetermined system, where infinitely many possible solutions can explain the observed data. This is particularly problematic when the number of measurements is very limited, as in the case of biplanar X-rays.

To achieve high-quality, artifact-free reconstructions, the Nyquist-Shannon sampling theorem dictates that dense sampling in the measurement space is essential [Shannon, 1949]. According to this theorem, the sampling rate must be at least twice the highest

frequency present in the signal to avoid information loss:

$$f_s \geq 2f_{\text{max}}. \tag{3.12}$$

In tomographic reconstruction, the "highest frequency" corresponds to spatial frequency, which represents the fine details of the object, while "sampling" refers to the angular distribution of projections. While hundreds of projections are sufficient for fine reconstruction, using only two projections results in a sampling rate far below the Nyquist limit. This causes significant information loss, making complete 3D reconstruction impossible and creating a highly ill-posed problem, where multiple solutions could fit the limited projection data.

To address sampling rates far below those suggested by the Nyquist-Shannon theorem, Compressed Sensing (CS) [Donoho, 2006] was developed.

In compressed sensing, the model is:

$$y = Ax + \eta, \tag{3.13}$$

where the number of measurements m is much smaller than the signal dimension n (i.e., $m \ll n$), making the system underdetermined. To enable effective recovery in this challenging scenario, compressed sensing assumes that x is sparse or approximately sparse in a transformed domain, such as the discrete cosine transform (DCT), wavelet, or Fourier basis [Candès, 2006; Donoho, 2006]. This sparsity, often further enhanced through total variation (TV) regularization, allows x to be accurately reconstructed from significantly fewer measurements than traditional methods typically require.

Reconstruction can be achieved using optimization techniques such as Basis Pursuit Denoising [Donoho, 2006] or LASSO (Least Absolute Shrinkage and Selection Operator) [Tibshirani, 1996]. In this approach, the signal x is represented as $x = \Psi\alpha$, where α is sparse in the transformed domain defined by Ψ.

- Basis Pursuit Denoising (BPD) minimizes the L_1-norm of the coefficient vector α while enforcing an exact fit to the measurements, as follows:

$$\min_{\alpha} \|\alpha\|_1 \quad \text{subject to} \quad y = A\Psi\alpha \tag{3.14}$$

  Here, BPD seeks the sparsest solution that perfectly matches the measurements. This is particularly suitable when measurements are noise-free or have very little noise, and an exact sparse solution is desired.

- LASSO, on the other hand, introduces a trade-off parameter λ that balances data fidelity with sparsity, making it more flexible and robust to noisy measurements. LASSO solves the following optimization problem:

$$\min_{\alpha} \frac{1}{2}\|y - A\Psi\alpha\|_2^2 + \lambda\|\alpha\|_1, \tag{3.15}$$

where λ controls the balance between fitting the observed data and enforcing sparsity in α.

In practice, this means that instead of requiring the signal x to be sparse in the spatial domain, it only needs to be sparse when represented in the transformed domain Ψx. For example, MRI often uses the Fourier basis as the sparsifying transform, leveraging the fact that most MR images are sparse in the frequency domain. Natural images, on the other hand, may use wavelet transforms, since natural images tend to have sparse representations in the wavelet domain.

Compressed sensing has enabled significant reductions in the number of measurements needed for successful reconstruction in various imaging inverse problems. Notably, it has shortened scan times in MRI, thereby accelerating clinical workflows [Lustig, 2007]. Despite these successes, sparsity-based methods have limitations when measurements are extremely sparse, as the sparsity assumptions are often hand-crafted or rely on simple learned sparse codes. Additionally, they may struggle when the signal does not naturally exhibit sparsity in a known basis [Bora, 2017].

To overcome these limitations, advanced methods like deep generative models have been explored as priors in compressed sensing. These models learn complex data distributions and structures beyond sparsity, providing strong priors that reduce ambiguity in the solution space—something classical methods lack, making them insufficient for accurate reconstructions.

3.2.3 Deep Learning Approaches for Inverse Problems in Imaging

In recent years, deep learning techniques have emerged as powerful data-driven methods for solving ill-posed inverse problems in imaging. These methods can be broadly categorized into two families: supervised approaches and distribution-learning approaches [Jin, 2017; Ongie, 2020].

- Supervised Methods: These techniques use large training datasets of measured images to learn the inverse mapping from measurements to images. Models like CNNs are trained for specific imaging tasks, capturing the relationship between measurements and the underlying structure to recover. While they perform well on in-domain data, their effectiveness tends to diminish significantly when faced with out-of-distribution scenarios, such as variations in anatomy or measurement noise [Ongie, 2020]. Additionally, these models suffer from averaging effects [Menon, 2020], which can lead to blurred outputs in reconstructions.

- Distribution-Learning Approaches: Unsupervised deep generative models, such as GANs and VAEs, provide a more flexible framework by learning the statistical distribution of plausible solutions. These models introduce stronger priors, ensuring that the reconstructed data lies directly on a learned manifold. As a result, they

can produce more robust and accurate reconstructions, even from sparse measurements [Ongie, 2020].

Feedforward Approaches

A lot of feedforward pipelines have been developed for a variety of tasks, including denoising, upsampling, and deblurring, by directly inverting measurements into target outputs. For instance:

- Super-Resolution (SR): Methods like SRGAN [Ledig, 2017] and EDSR [Lim, 2017] reconstruct high-resolution images from low-resolution inputs, significantly outperforming classical techniques like bilinear or bicubic interpolation by producing sharper and more detailed results. However, these models can still introduce artifacts or lose fine texture due to the averaging effect, as they tend to blend between solutions that downsample effectively [Menon, 2020].
- Denoising: CNN-based methods like DnCNN [Zhang, 2017] effectively remove noise from images by learning a direct mapping from noisy inputs to clean outputs. While these models demonstrate impressive performance on benchmark datasets, their generalization to unseen noise types or levels remains a challenge.
- Deblurring: GAN-based models such as DeblurGAN [Kupyn, 2018] and its successor DeblurGAN-v2 [Kupyn, 2019] have shown success in recovering sharp images from blurry inputs. These models leverage adversarial training to produce sharper reconstructions than traditional methods, though they may still struggle with severe blur or complex image structures.
- Inpainting: Inpainting methods, such as DeepFill [Yu, 2018] and EdgeConnect [Nazeri, 2019], fill in missing or corrupted parts of an image by learning to predict the missing content from the surrounding context. These models are particularly useful for tasks like image restoration, object removal, and filling in gaps in images. By leveraging generative networks, they can produce visually plausible inpainted regions, but they may struggle with complex structures or fine details, especially when large areas are missing [Marinescu, 2020].

This direct learned inversion approach is widely used in the literature to tackle imaging inverse problems.
Recent advances in combining compressed sensing with generative models have demonstrated improved robustness and superior results compared to traditional approaches, particularly when dealing with sparse data or complex perturbations in measurements.

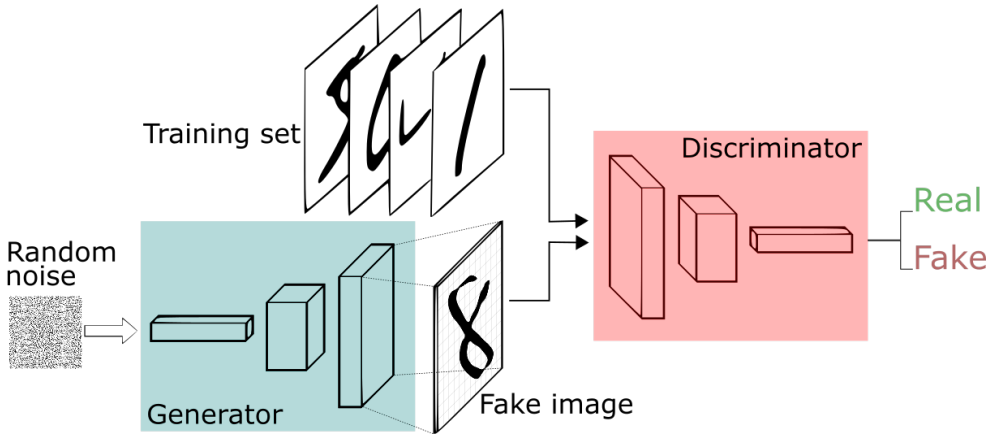

Figure 3.9: Overview of a GAN [Goodfellow, 2014]. The generator creates a "fake" image from random noise, attempting to resemble images from the training set. The discriminator evaluates both real and fake images, classifying them as "real" or "fake." Through adversarial training, both components improve, leading to realistic image generation.

Compressed Sensing with Generative Models

Network architectures that integrate both CNN-based blocks and the imaging forward model have become increasingly popular, as they combine deep learning to model complex data distributions and the mathematical framework of compressed sensing. Recent advances in deep generative models have transformed compressed sensing, offering a powerful alternative to traditional sparsity-based methods, allowing to recover signals from incomplete or noisy measurements [Hammernik, 2018; Aggarwal, 2018; Mardani, 2018]

Compressed Sensing Using Generative Models (CSGM), introduced by [Bora, 2017], leverages the ability of generative models to approximate the true distribution of complex signals, such as images, by representing them in a more structured and compact form—specifically, in a low-dimensional latent space. These generative models are trained to map low-dimensional latent codes $z \in \mathbb{R}^k$ to high-dimensional signals $x = G(z)$, where $G : \mathbb{R}^k \to \mathbb{R}^n$ is the generator function that reconstructs signals from the latent space. CSGM uses these pre-trained generative models as priors, allowing them to approximate the statistical properties of the data and produce high-quality reconstructions, even from sparse measurements. These generative models are highly effective at capturing complex image statistics, which makes them particularly suited for solving a variety of inverse problems.

Generative models Generative models are a class of machine learning models designed to learn the underlying distribution of a dataset and generate new data points that are similar to the original data. These models aim to model the joint probability distribution $p(x, z)$, where x is the data (e.g., images, text) and z is a latent representation. By learning this distribution, the model can generate new samples by sampling from the learned latent space and decoding these samples back into data space using the generator

function $G(z)$.

Four main types of generative models are commonly used in compressed sensing applications:

- Variational Autoencoders (VAEs)[Kingma, 2013]: VAEs are probabilistic generative models that learn a latent representation z of the data x by maximizing a variational lower bound (ELBO) on the data likelihood $\log p(x)$. This is done by introducing a variational distribution $q(z|x)$, typically Gaussian, to approximate the true posterior $p(z|x)$. The model consists of an encoder network $q_\phi(z|x)$ that maps the data x to the latent code z, and a decoder network $p_\theta(x|z)$ that generates reconstructed data from the latent variable. The objective is to minimize the Kullback-Leibler divergence $KL(q_\phi(z|x) \parallel p_\theta(z))$, where $p_\theta(z)$ is the prior on the latent space (often an isotropic Gaussian), along with the reconstruction error $\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)]$. This regularization helps ensure smooth transitions between generated samples in the latent space and allows VAEs to perform efficient sampling and data reconstruction.

- Generative Adversarial Networks (GANs)[Goodfellow, 2014]: GANs (illustrated in Figure 3.9) consist of two networks—a generator $G_\theta(z)$ and a discriminator $D_\phi(x)$—that are trained in a minimax game. The generator maps a latent variable z, sampled from a simple prior distribution $p(z)$ (typically a standard normal distribution), to a data sample $G_\theta(z)$. The discriminator $D_\phi(x)$ estimates the probability that a given sample is real (from the data distribution $p_{\text{data}}(x)$) or fake (generated by $G_\theta(z)$). The training objective is to solve the following optimization problem: $\min_G \max_D \mathbb{E}_{x\sim p_{\text{data}}}[\log D_\phi(x)]+\mathbb{E}_{z\sim p(z)}[\log(1-D_\phi(G_\theta(z)))]$. The generator aims to minimize this objective by fooling the discriminator, while the discriminator tries to maximize it by distinguishing real from fake samples. GANs are capable of producing high-quality samples, but their training can be unstable due to issues like mode collapse and the adversarial nature of the optimization.

- Flow-based Models[Dinh, 2014]: Flow-based models explicitly model the data distribution $p(x)$ by learning an invertible mapping f_θ between the data space x and a latent space z, where z follows a simple prior distribution (e.g., a Gaussian). This mapping is bijective, meaning that both forward and inverse transformations $x \leftrightarrow z$ can be computed exactly. The data likelihood $p(x)$ is computed using the change of variables formula: $p(x) = p_z(f_\theta(x))|\det(J_{f_\theta}(x))|$, where $p_z(z)$ is the prior distribution in the latent space, and $J_{f_\theta}(x)$ is the Jacobian determinant of the transformation f_θ. Flow-based models are trained by maximizing the exact log-likelihood $\log p(x)$, and their invertible structure makes them particularly useful for tasks requiring exact likelihood estimation and sampling. Normalizing flows, a common class of flow-based models, utilize a series of simple, invertible transformations to model complex data distributions in a computationally efficient manner, which is advantageous for compressed sensing tasks.

- Diffusion Models[Sohl-Dickstein, 2015]: recently developed, diffusion models generate data by gradually reversing a stochastic process that corrupts data into noise over multiple time steps, modeled as a Markov chain. In the forward process, noise is added step-by-step to the data x_0 to create a sequence of increasingly noisy samples $x_1, x_2, \ldots, x_T$, where x_T is nearly pure Gaussian noise. This forward process is typically a fixed Markov process, where each x_t is conditioned only on the previous step, x_{t-1}, and noise is added according to a Gaussian distribution. The reverse process, parameterized by a neural network, learns to predict $p(x_{t-1}|x_t)$, the conditional probability of the previous state given the current noisy state. The model is trained using denoising score matching to estimate the noise at each step. By iteratively denoising, the model generates high-quality samples starting from pure noise x_T and refining it to x_0. Diffusion models are particularly powerful for compressed sensing applications due to their ability to recover data from incomplete or noisy measurements while providing a tractable likelihood function $p(x_0)$, making them well-suited for both probabilistic modeling and data recovery tasks.

  GANs are typically used for their ability to generate high-quality, realistic samples, making them valuable in compressed sensing for tasks that require sharp, detailed reconstructions. For example, the introductory paper [Bora, 2017] employs a Deep Convolutional GAN (DCGAN) [Radford, 2015]. However, diffusion models and VAEs offer robust probabilistic modeling, which can be advantageous in handling noise and uncertainty. Flow-based models, with their exact likelihood estimation, provide a useful alternative when invertibility and precise probability control are essential for reconstruction tasks. Together, these models cover a range of capabilities in compressed sensing applications, addressing different aspects of data quality, flexibility, and computational efficiency.

Compressed Sensing with Generative Model The task in CSGM is to find the latent vector $z \in \mathbb{R}^k$ such that the corresponding signal $G(z)$, generated by the model, minimizes the measurement error with respect to the observed data y. This can be expressed by the following optimization problem:

$$\min_{z \in \mathbb{R}^k} \|AG(z) - y\|_2^2, \tag{3.16}$$

where A is the measurement matrix, and y represents the measurements. Since the generative model imposes strong priors on z, the solution space is constrained, allowing the model to operate effectively even when the data is highly undersampled.

Bora et al. extend this formulation by adding a regularization term $L(z)$ to further guide the optimization. This term encourages the optimization process to remain within regions of the latent space that correspond to plausible solutions, as preferred by the generative model. This regularizer often takes the form of an ℓ_2-norm, which aligns with the isotropic Gaussian prior typically imposed on the latent variable z. The full regularized

objective function becomes:

$$\min_{z \in \mathbb{R}^k} \|AG(z) - y\|_2^2 + \lambda \|z\|_2^2, \tag{3.17}$$

where λ balances the measurement error with the strength of the prior. This regularization encourages z to remain close to the prior distribution, ensuring that the reconstruction is not only consistent with the measurements but also plausible under the learned generative model.

A significant challenge in CSGM is that the optimization is non-convex, primarily due to the generator $G(z)$, which is typically modeled as a deep neural network with multiple non-linear layers. These non-linearities create a complex optimization landscape with numerous local minima and saddle points, making it difficult to converge to the global minimum. Additionally, the ill-posed nature of inverse problems often leads to multiple possible solutions that fit the data, further complicating the optimization process.

However, Bora et al. demonstrated empirically that gradient-based methods, such as stochastic gradient descent (SGD), are often able to find solutions that are close to the true signal. This is largely due to the strong priors imposed by the generative model, which significantly reduce the search space, making it easier to avoid poor local minima. The generative model effectively guides the optimization toward regions of the latent space that correspond to realistic reconstructions, thereby being closed to the solution even in the presence of non-convexity.

In addition to these empirical findings, Bora et al. provided interesting theoretical guarantees for CSGM. They showed that if the generative model G is L-Lipschitz, then the number of random Gaussian measurements required to recover the signal x with a small error grows only as $O(k \log n)$, where k is the dimensionality of the latent space. This result is a major improvement over traditional compressed sensing, where the number of measurements typically scales with the signal dimension n. By restricting the solution to the range of the generator, CSGM achieves a significant reduction in the required number of measurements.

The recovery guarantee relies on a generalization of the Restricted Eigenvalue Condition (REC), known as the Set-Restricted Eigenvalue Condition (S-REC) [Bora, 2017]. This condition ensures that the difference between any two vectors in the range of the generator is well-separated in the measurement space. Formally, for a measurement matrix A, the S-REC is satisfied for a set $S \subset \mathbb{R}^n$ if for all $x_1, x_2 \in S$,

$$\|A(x_1 - x_2)\|_2 \geq \alpha \|x_1 - x_2\|_2 \tag{3.18}$$

for some constant $\alpha > 0$. This ensures that the optimization procedure can accurately recover the latent code z such that the generated signal $G(z)$ is close to the true signal x. Bora et al. proved that random Gaussian measurement matrices satisfy the S-REC with high probability for the range of commonly used generative models, such as VAEs and GANs, ensuring robust recovery even with a reduced number of measurements.

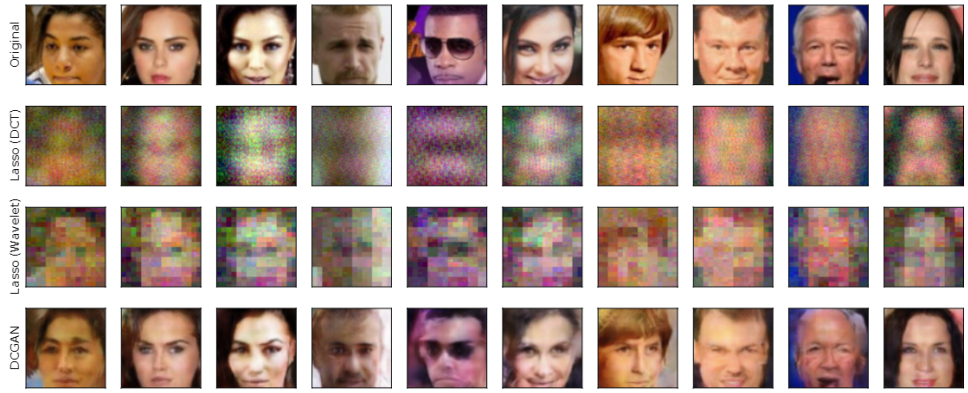

Figure 3.10: Reconstruction results on CelebA with m = 500 measurements (of n = 12288 dimensional vector). It shows original images (top row), and reconstructions by Lasso with DCT basis (second row), Lasso with wavelet basis (third row), and CSGM with DCGAN (last row)[Bora, 2017].

Bora et al. showed empirically that CSGM can outperform traditional compressed sensing methods like LASSO on various image datasets. They focused on the task of super-resolution, which involves constructing a high-resolution image from a low-resolution version of the same image. This can be viewed as a special case of the general framework of inverse problem with linear measurements, where the measurements represent local spatial averages of the pixel values. For example, in their experiments on the MNIST dataset [LeCun, 1998], CSGM, using a DCGAN [Radford, 2015], was able to reconstruct images with high accuracy using only 25 measurements, while LASSO required around 400 measurements to achieve similar performance.

Figure 3.10 presents a comparison of reconstruction results on the CelebA dataset [Liu, 2015b] using Lasso with a DCT basis, Lasso with a wavelet basis, and CSGM. The CSGM approach demonstrates significantly better results with fewer measurements compared to the sparsity-based methods.

They show that the total reconstruction error in CSGM can actually be decomposed into three components:

- Representation error: The discrepancy between the true signal and the closest signal in the range of the generator.
- Measurement error: The error due to the finite number of measurements and noise in the observations.
- Optimization error: The error due to the optimization process not finding the global minimum of the objective function.

Since the generative model is only an approximation of the true data distribution, there may be some signals that cannot be perfectly represented by the generator. This error, called the representation error, arises when the true signal x^* lies outside the range

of the generator. In such cases, even with perfect measurements, the recovery process will not yield an exact reconstruction of x^*.

In practice, the representation error is often the dominant source of error, especially when the generative model is not expressive enough. In an experiment, they tested image reconstruction within the range of the generator by sampling a latent vector z^*, generating a signal $x^* = G(z^*)$, and treating it as a real image. This eliminated representation error, allowing them to focus on evaluating the optimization process. Results showed near-perfect reconstructions with few measurements, indicating that the optimization effectively minimized the objective, with both optimization and measurement errors being small. This demonstrates that, with well-trained or more expressive generative models that capture the solution space accurately, high-quality reconstructions can be achieved with far fewer measurements.

CSGM has been successfully applied to numerous inverse problems, such as non-linear phase retrieval [Bahmani, 2017], and has been further improved with techniques like invertible models [Kruse, 2021], sparsity-based deviations [Lustig, 2008], image adaptivity [Ulyanov, 2018], and posterior sampling [Hoffman, 2017]. These advancements have enhanced the robustness and performance of CSGM across a wide range of applications.

To further progress in complex inverse problems, sophisticated generative models like BigGAN [Brock, 2019] and StyleGAN [Karras, 2019; Karras, 2020b] have set new standards in image quality, scalability, and control. BigGAN enables high-resolution reconstructions, while StyleGAN models offer fine control over image attributes, making them valuable for tasks like inversion and compressed sensing where precision is essential.

Style-Based Generative Models StyleGAN, introduced by [Karras, 2019], is a groundbreaking model in generative networks, known for generating high-resolution images with fine control over semantic and stylistic attributes. It revolutionized the generator architecture by introducing a more controllable and disentangled latent space, enabling smoother transitions in generated content and offering users the ability to manipulate specific image features. StyleGAN's architectural innovations result in superior performance in high-resolution image generation and diversity. Examples of generation are presented in 3.11

Figure 3.12 presents the architecture. The generator consists of two main components: a mapping network and a synthesis network. The mapping network transforms the latent vector $z \in \mathbb{R}^{512}$, sampled from a Gaussian normal distribution, into an intermediate vector $w \in \mathbb{R}^{512}$. This intermediate latent space $\mathcal{W}$ controls the synthesis process. The synthesis network, consisting of 18 layers, progressively increases the image resolution, doubling it at each step from 4×4 to 1024×1024. The final layer outputs an RGB image using a 1×1 convolution.

Unlike traditional models that feed the latent code directly into the input layer [Brock, 2019], StyleGAN's innovation lies in mapping the latent space $Z \in \mathbb{R}^{512}$ to $\mathcal{W}$ using a non-linear 8-layer MLP. This intermediate space $\mathcal{W}$ provides a more structured and

Figure 3.11: Examples of diverse artificial face images generated by StyleGAN using the Flickr-Faces-HQ dataset, a high-quality face image dataset [Karras, 2019].

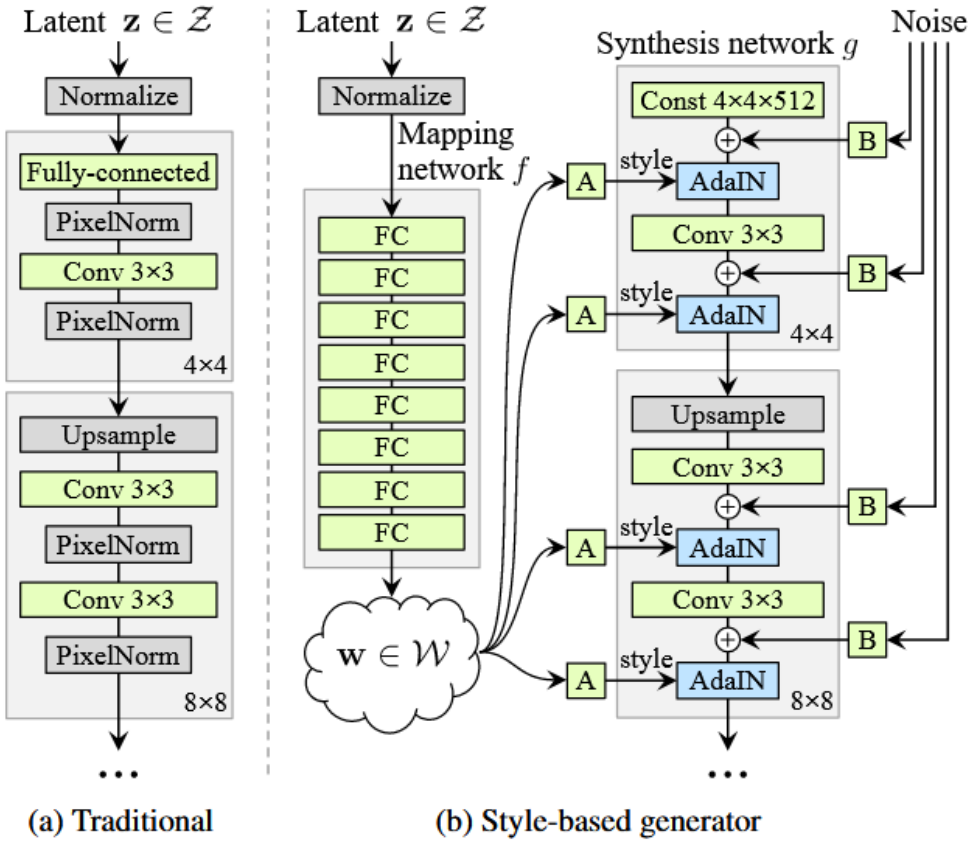

Figure 3.12: Unlike traditional generators that feed the latent code directly into the input layer, StyleGAN first maps the latent code to an intermediate latent space $\mathcal{W}$. This space then controls the synthesis network through AdaIN at each convolutional layer, providing more control over the image generation process. Noise is added at each convolutional layer before applying nonlinearity. "A" represents a learned affine transform, and "B" applies per-channel scaling factors to the noise input. The synthesis network progressively increases resolution from 4^2 to 1024^2, with the final output converted to RGB using a 1×1 convolution [Karras, 2019].

disentangled representation, allowing for precise control over the image generation process. Each layer of the synthesis network is modulated through adaptive instance normalization (AdaIN) [Huang, 2017], which adjusts the mean and variance of feature maps, giving users the ability to manipulate attributes like color, texture, and style at different levels.

Learned affine transformations specialize the vector $\mathcal{W}$ into *styles* $y = (y_s, y_b)$, which control the AdaIN operations after each convolutional layer in the synthesis network. The AdaIN operation is defined as:

$$\text{AdaIN}(x_i, y) = y_{s,i} \cdot \frac{x_i - \mu(x_i)}{\sigma(x_i)} + y_{b,i}. \tag{3.19}$$

Here, each feature map x_i is normalized independently, and the style vector y controls the scale and bias applied to these feature maps. The dimensionality of y is twice the number of feature maps at each layer. In addition to style control, StyleGAN introduces

explicit noise inputs: single-channel Gaussian noise images are injected into each layer of the synthesis network. These noise inputs are scaled and added to the feature maps, introducing stochastic variations and allowing the model to generate more highly detailed and realistic images. The combination of noise and latent space control ensures that StyleGAN produces fine-grained details while maintaining high-level structural consistency.

A key element of StyleGAN is *style mixing*, which allows mixing coarse and fine styles during both training and inference. This technique enables flexible and controlled image generation by combining different styles at various levels of resolution, such as blending overall structure with finer details, leading to richer and more diverse outputs, as shown in Figure 3.13,

Also, StyleGAN's unique latent space, $\mathcal{W}$, is a key feature that enables smooth transitions and precise control over generated content, setting it apart from models like BigGAN, which emphasize high-quality synthesis but lack the same level of semantic control. StyleGAN's mapping network enhances the disentanglement of $\mathcal{W}$, making factors such as pose, color, and texture more linearly separable. This design allows for intuitive manipulation of specific attributes—like facial expressions or lighting—while preserving the coherence of generated images. Additionally, $\mathcal{W}$ enables smooth transitions, facilitating gradual transformations in features such as age, hairstyle, or background when interpolating between images. This is achieved by optimizing the introduced *perceptual path length*, which measures the smoothness of image changes within the latent space, allowing for finer control and realistic edits by disentengling attributes more effectively.

StyleGAN2 [Karras, 2020b] improved the original StyleGAN by eliminating artifacts through a redesigned generator architecture, removing AdaIN in favor of a modulation/demodulation mechanism, and introducing better path length regularization for smoother transitions in the latent space. It also enhanced feature disentanglement, allowing for more precise control of individual attributes, and improved fine detail synthesis, resulting in cleaner, more realistic images with better high-frequency detail handling.

In summary, style-based generative models represent a significant leap in generative modeling, offering high-resolution image generation, fine control, and detailed stochastic features. These capabilities are especially valuable for inverse problems, compressed sensing, and semantic transformations, with practical applications in areas like image-to-image translation, domain adaptation, and tomographic imaging [Bhadra, 2022], which we will explore next.

Figure 3.13: Style Mixing. Images generated from two latent codes (sources A and B) are shown, with mixed styles applied. Copying coarse styles (4^2 to 8^2) from source B transfers high-level features like pose and face shape, while fine details remain from A. Mid-level styles (16^2 to 32^2) from B affect smaller features like hairstyle and eye state, while overall shape from A is preserved. Fine styles (64^2 to 1024^2) from B mainly alter colors and textures [Karras, 2019].

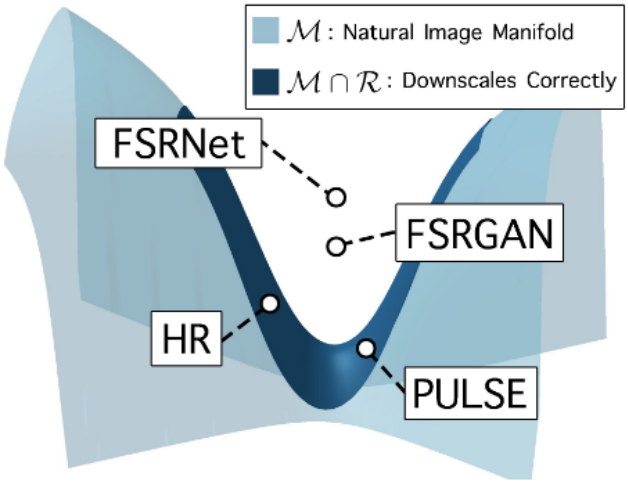

Figure 3.14: Feedforward methods, like FSRNet [Chen, 2018], often produce averaged images that downscale accurately. Although adding a discriminator loss in FSRGAN [Wang, 2018b] helps guide outputs toward the natural image manifold, it doesn't fully resolve the issue. In contrast, PULSE [Menon, 2020] consistently moves along the manifold, producing high-quality results.

CSGM for High-Quality Image Upsampling A relevant example of applying more advanced generative models with CSGM is Photo Upsampling via Latent Space Exploration (PULSE) [Menon, 2020], which presents a novel approach to super-resolution by leveraging the latent space of state-of-the-art generative models, such as StyleGAN2 [Karras, 2020b], to achieve high-quality image reconstructions. Following the previous trend, PULSE by exploring the latent space of generative models to find realistic high-resolution images that correctly downscale to the low-resolution input.

As with previous compressed sensing approaches using unsupervised generative models, one of PULSE's key advantages is its self-supervised nature. This eliminates the need for paired low- and high-resolution datasets required by feedforward methods. Instead, PULSE operates without relying on specific degradation models, enabling it to generalize across different scenarios without the need for retraining. The only drawback is that the corruption process must be known in advance.

Menon et al. observe that traditional super-resolution methods typically minimize pixel-wise loss functions, which leads to an undesirable *averaging effect*. Indeed many solutions can downscale well to the input images. Traditional method minimize the loss by averaging all solutions that downscale well, which results in blurred details. Even methods that try to integrate GAN, can tend the solutions to be closer to the realistic manifold, but it is still an averaging of pixelwise- and GAN-based solutions. PULSE shows that it overcomes this issue by directly navigating the latent space of a generative model to find solutions that lie on the natural image manifold. This guarantees that the super-resolved images are well realistic and maintain high perceptual quality. Figure 3.14 illustrates this point.

Menon et al. note that simply ensuring $z \in L$ doesn't guarantee that $G(z) \in M$, the manifold of realistic images. A common solution is to impose a prior on L and add a

regularization term, such as $l2$ regularization for Gaussian priors [Bora, 2017]. However, this tends to push vectors towards the origin, while most of the probability mass in high-dimensional Gaussian distributions lies near the surface of a sphere with radius $\sqrt{d}$, a *soap effect*.

To address this, PULSE replaces the Gaussian prior with a uniform prior on the hypersphere surface, where $\|z\|_2 = \sqrt{d}$ (with z as the latent vector and d as the dimensionality). This constraint ensures realistic image generation and more efficient search by minimizing:

$$\min_{z \in \sqrt{d} S^{d-1}} \|DS(G(z)) - I_{LR}\|_p, \tag{3.20}$$

where DS is the downscaling operator, $G(z)$ is the generator, and I_{LR} is the low-resolution input. This approach maintains a balance between realistic generation and matching low-resolution data, starting from a random latent vector initialization.

Optimizing directly in the latent space $z \in S_{512} \subset \mathbb{R}^{512}$ often leads to poor results due to limited expressiveness. Instead, using the full 18×512-dimensional latent space, denoted as $\mathcal{W}^+$, allows more flexible inputs that better match the synthesis network. However, allowing the 18 vectors $w_1, w_2, \ldots, w_{18}$ to vary independently risks deviating from the natural image manifold.

To balance flexibility with realism, a geodesic cross loss term Geocross—similar to negative cosine similarity loss—is introduced, penalizing large angular deviations between input vectors:

$$\text{Geocross}(w_1, \ldots, w_k) = \sum_{i<j} \theta(w_i, w_j)^2. \tag{3.21}$$

This constraint enhances expressiveness while keeping the generated images close to the natural image manifold, resulting in more realistic outputs.

One advantage of PULSE over traditional feedforward methods is its ability to produce multiple plausible high-resolution outputs for the same low-resolution input. By using projected gradient descent with Adam [Kingma, 2014] and random initialization, PULSE explores different local minima due to the non-convexity of the optimization problem, producing a diverse set of high-resolution, photo-realistic images that vary in details yet remain consistent with the low-resolution input.

An additional benefit is PULSE's robustness against degradations like noise and motion blur. Rather than directly matching the degraded input, PULSE projects outputs onto the realistic image manifold, ensuring they downscale accurately to the true, non-degraded low-resolution image. This contrasts with traditional supervised models, which require explicit training on noisy data to handle degradation, whereas PULSE achieves it without extra training.

Extensive experiments demonstrated PULSE's ability to outperform state-of-the-art feedforward methods, producing perceptually superior images that preserve fine details, such as facial features. By effectively navigating the latent space of generative models, PULSE achieves high-quality reconstructions with fewer artifacts and blurring issues than

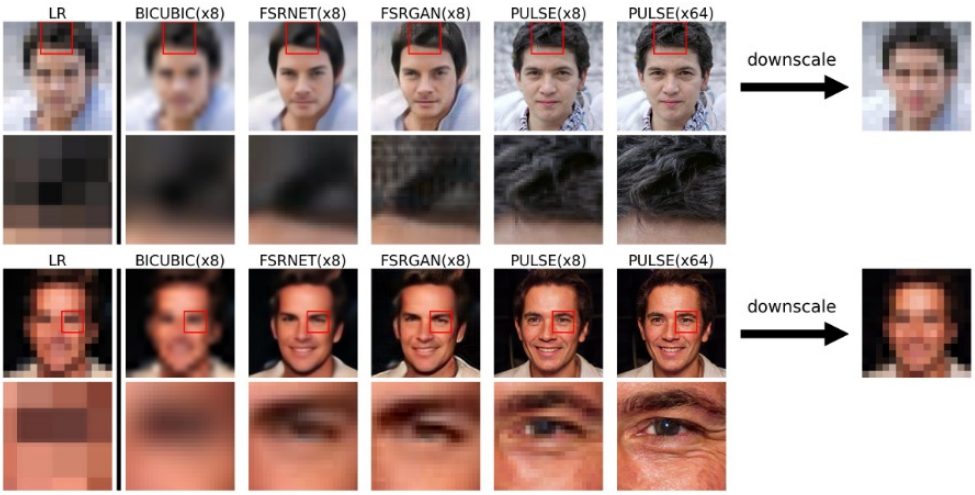

Figure 3.15: Comparison of PULSE with bicubic upscaling, and feedforward methods FSRNet and FSRGAN. Feedforward methods often suffer from blurring and averaging effects, while PULSE produces more realistic and higher-quality results [Menon, 2020]

feedforward methods, as shown in Figure 3.15.

A Bayesian Perspective Building on previous advancements like PULSE, which leverages latent space exploration for high-quality image upsampling, Bayesian Reconstruction using Generative Models (BRGM) [Marinescu, 2020] introduces a Bayesian framework for image restoration tasks such as super-resolution and inpainting, utilizing pre-trained generative models like StyleGAN2. Unlike PULSE, which operates through direct latent space navigation, BRGM derives its loss function from the Bayesian MAP estimate, employing variational inference to sample from the approximate posterior distribution. This Bayesian approach inspire our core method for 3D reconstruction with biplanar X-rays, detailed in Section 3.4.

The method formulates the problem as finding the latent vector w that maximizes the posterior probability of a clean image, given corrupted observation, combining both the generative prior and the corruption or forward process within a Bayesian framework.

The main goal of BRGM is to recover the optimal clean image I^*_{CLN} from a corrupted observation I. The corrupted image I is generated by applying a forward process f to the clean image, which is modeled by the generator $G(w)$, where w is the latent vector of the generator. The corruption model f includes processes like downsampling or masking:

$$I = f(G(w)) + \epsilon, \tag{3.22}$$

where ϵ represents noise or other distortions. The clean image is recovered by finding the MAP estimate of the latent vector w, which maximizes the posterior probability $p(w|I)$.

Using Bayes' theorem, this posterior is defined as:

$$p(w|I) \propto p(I|f \circ G(w))\, p(w), \tag{3.23}$$

where $p(w)$ is the prior distribution over the latent space, typically modeled as a Gaussian: $p(w) \sim \mathcal{N}(0, I)$ and $p(I|f \circ G(w))$ is the likelihood of observing I given the clean image generated by $G(w)$.

The MAP estimate for w is computed by maximizing this posterior:

$$w^* = \arg\max_w \; p(w)p(I|f \circ G(w)). \tag{3.24}$$

Once w^* is obtained, the clean image is reconstructed as:

$$I^*_{CLN} = G(w^*). \tag{3.25}$$

The generative model G acts as a prior over the clean image space, ensuring that the reconstructed image remains on a realistic image manifold. The prior term $p(w)$ is based on the structure of the generator's latent space. As in PULSE, they relax the constraint that all w_i must be equal, allowing each to be optimized independently, resulting in the extended latent space $\mathcal{W}^+$. The prior over the latent vectors w_i is modeled with two components:

- Gaussian Prior: Ensures the latent vectors w_i remain close to those observed during training:

$$L_w = \prod_{i=1}^{18} N(w_i|\mu, \sigma^2), \tag{3.26}$$

  where μ and σ^2 are the mean and variance of the latent vectors.

- Cosine Similarity Prior: similar to the Geocross loss introduced in [Menon, 2020], this forces alignment between latent vectors w_i and w_j for different layers, modeled using the von Mises distribution:

$$L_{colin} = \prod_{i,j} \mathcal{M}\left(\cos^{-1}\frac{w_i^T w_j}{|w_i||w_j|}\middle|0, \kappa\right), \tag{3.27}$$

  where κ controls the strength of the alignment.

The MAP estimate (Eq. 3.24) can be expressed as a weighted sum of four loss terms:

$$w^* = \arg\min_w \left(\lambda_w L_w + \lambda_c L_{\text{colin}} + \lambda_{\text{pixel}} L_{\text{pixel}} + \lambda_{\text{percept}} L_{\text{percept}}\right), \tag{3.28}$$

where L_w is the prior loss over w, L_{colin} is the colinearity loss to ensure alignment between latent vectors, L_{pixel} is the pixel-wise loss comparing the corrupted image to the generated one, and L_{percept} is the perceptual loss [Johnson, 2016], which compares high-level features between the corrupted and generated images.

In ill-posed problems with multiple solutions, a Bayesian framework is advantageous, as it captures a distribution of possible outcomes, unlike MAP, which provides only a single estimate and lacks the full range of plausible solutions. Sampling from the posterior $p(x|y)$ enables exploration of this solution space but is often computationally intensive due to complex priors and high dimensionality. Deep learning methods have shown promise in approximating the posterior by learning data structures from large datasets, effectively capturing complex priors without manual encoding. BRGM shows this capacity to sample directly from the estimated posterior. Unlike PULSE, which uses empirical sampling by restarting optimization from various initializations [Menon, 2020], BRGM employs variational inference [Hinton, 1993; Graves, 2011] to approximate $p(w|I)$ and generate multiple plausible reconstructions.

The goal is to find a parametric distribution $q(w|\theta)$, where θ represents the learnable parameters, to approximate the true posterior. This is achieved by minimizing the Kullback-Leibler (KL) divergence between $q(w|\theta)$ and $p(w|I)$:

$$\theta^* = \arg\min_\theta \mathrm{KL}[q(w|\theta)\|p(w|I)] = \arg\min_\theta \int q(w|\theta) \log \frac{q(w|\theta)}{p(w)p(I|w)}\, dw. \tag{3.29}$$

The objective is to minimize the expected value over $q(w|\theta)$ by approximating it with Monte Carlo samples. The expected value is approximated by taking Monte Carlo samples $w^{(i)}$ from $q(w|\theta)$. The objective becomes:

$$\theta^* = \arg\min_\theta \sum_{i=1}^{n} \left(\log q(w^{(i)}|\theta) - \log p(w^{(i)}) - \log p(I|w^{(i)}) \right), \tag{3.30}$$

where $w^{(i)}$ are the Monte Carlo samples drawn from $q(w|\theta)$.

The distribution $q(w|\theta)$ is parameterized as a Gaussian. The Gaussian is sampled using unit noise ϵ, shifted by the variational mean μ_v, and the transformed standard deviation σ_v, which is re-parameterized as:

$$\sigma_v = \log(1 + \exp(\rho_v)), \tag{3.31}$$

where $\theta = [\mu_v, \rho_v]$ are the variational parameters.

The final objective function to optimize becomes:

$$\theta^* = \arg\min_\theta \left(-\log p(\theta) + \sum_{i=1}^{n} \left(\log q(w^{(i)}|\theta) - \log p(w^{(i)}) - \log p(I|w^{(i)}) \right) \right), \tag{3.32}$$

where the prior $p(\theta)$ is modeled as an inverse gamma distribution over σ_v, which encourages larger standard deviations.

Figure 3.16 demonstrates generation sampling using variational inference, presenting diverse solutions that align with the measurements.

BRGM outperformed other super-resolution methods, particularly at lower input res-

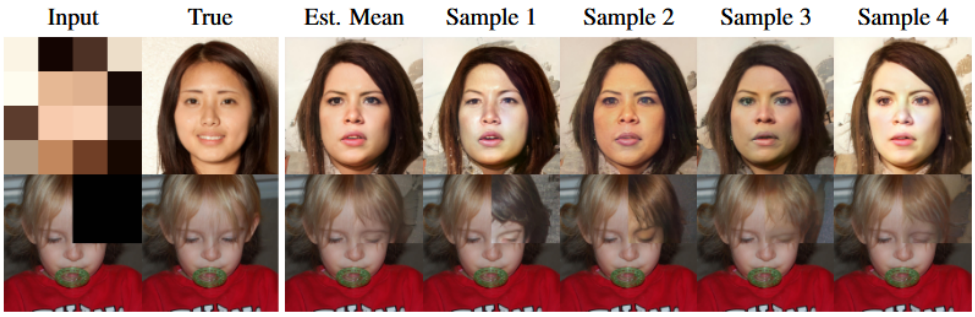

Figure 3.16: Generation sampling using variational inference from [Marinescu, 2020]. The figure shows the input image (left column), the target (second column), the estimated mean image $G(\mu_v)$ (third column), and samples generated around the mean $G(\mu_v + \sigma_v \epsilon)$ (right column).

olutions. Feedforward methods like ESRGAN [Wang, 2018a] produced jittery artifacts, while SRFBN [Li, 2019] resulted in overly smooth images. While PULSE generated realistic high-res outputs, they often didn't match the true image due to strict projection onto the unit sphere. BRGM, by relaxing this constraint with a soft prior L_w, allowed for more accurate reconstructions. Both BRGM and PULSE can achieve super-resolution beyond 4 times, up to 1024×1024, without dataset-specific tuning.

While BRGM produced diverse and high-quality reconstructions through variational inference, it sometimes overfitted the posterior and had difficulty generalizing to unseen images, a common issue with learning-based methods. Furthermore, over-representation of certain populations in the training data can introduce biases, leading to inaccurate reconstructions. These biases can cause discrepancies between downsampled images and the original inputs in both super-resolution and inpainting tasks

GAN Inversion All the previous methods rely on what is generally known as GAN inversion. GAN inversion, which explores the latent space of learned distributions with GAN, has become a powerful tool for image restoration and 3D reconstruction. By projecting images into the GAN's latent space, this method enables applications like upsampling [Menon, 2020], inpainting [Li, 2020], noise reduction [Bau, 2018], and artifact removal [Karras, 2019]. Leveraging these learned priors allows for tasks such as super-resolution [Menon, 2020], 3D shape reconstruction [Wu, 2015; Nguyen-Phuoc, 2019], image completion [Li, 2020], and stylization [Karras, 2019], demonstrating GAN inversion's broad utility for inverse problems.

Compressed Sensing for Medical Imaging

Beyond traditional computer vision, compressed sensing with generative models has become increasingly valuable in medical imaging, where the goal is to reconstruct high-quality images from incomplete or noisy patient data. Generative models have shown great promise in medical imaging for tasks such as MRI reconstruction, tomographic imaging, and solving inverse problems for medical imaging.

For example, AmbientFlow [Kelkar, 2023a] leverages invertible generative models to reconstruct images from incomplete and noisy data. For accelerated MRI, score-based diffusion models have been shown to significantly improve reconstruction quality, as described by Chung et al. [Chung, 2022]. Shang et al. [Shang, 2024] introduced ResDiff, a model combining CNNs and diffusion for image super-resolution. Bayesian imaging with neural network priors, explored by Holden et al. [Holden, 2022], has improved uncertainty modeling in reconstructions. StyleGAN has also been adapted for medical imaging, with style-based generative models used for image-constrained reconstruction, as presented by Kelkar and Anastasio [Kelkar, 2021b]. GANs have been employed to learn canonical medical image statistics, as demonstrated by Kelkar et al. [Kelkar, 2023b], while score-guided diffusion models have been applied to fast Langevin mixing for inverse problems, as seen in Daras et al. [Daras, 2022].

Robust Compressed Sensing for MRI The CSGM framework [Bora, 2017] has shown that deep generative priors can be powerful tools for solving inverse problems. However, before [Jalal, 2021a] this framework has been empirically successful only on certain natural datasets (for example, human faces and MNIST digits), and it is known to perform poorly on real out-of-distribution samples.

Jalal et al. presented the first successful application of the CSGM framework to clinical MRI data, using a score-based generative model trained on the fastMRI dataset. The method, which employs posterior sampling via Langevin dynamics, achieves high-quality reconstructions and remains robust to changes in data distribution and measurement processes.

Without assuming a specific measurement system, the model reconstructs undersampled MRI data across various sampling schemes, showing competitive performance with end-to-end models on in-distribution data and superior robustness to out-of-distribution shifts. Theoretical results support that posterior sampling is near-optimal, even with imperfect priors.

Additionally, the method allows for uncertainty quantification by generating multiple reconstructions with different random initializations. This flexibility and robustness make it a strong candidate for clinical use, as it can handle variations in sampling, hardware, and anatomy in real-world MRI settings.

Compressed Sensing for Sparse Tomography Let's now return to the ill-posed problem of tomographic reconstruction with sparse measurements, using only a few projections.

Alongside our development of a method for 3D reconstruction from biplanar X-rays, a related approach emerged as the first to use compressed sensing with generative models for tomographic imaging with sparse projections. This approach, PULSE++ [Bhadra, 2022], extends the original PULSE methodology to tomographic imaging, marking the first adaptation of PULSE—originally developed for super-resolution—to tackle ill-posed 2D tomographic reconstruction challenges.

Few rectifications have been made compared to PULSE to account for the quality requirements and specificity of tomographic imaging. The assumption that the latent vectors in StyleGAN's latent space $\mathcal{W}^+$ follow a Gaussian structure, as used in the PULSE method, was rigorously evaluated and found to be inaccurate. The projection of latent vectors onto the spherical surface $S_{k-1}(\sqrt{k})$, as assumed by PULSE, increases the risk of data inconsistency. In response to these findings, PULSE++ replaces the projection onto the spherical surface with a projection onto an annular region $A = \{w_i \in \mathbb{R}^k \mid \delta_{\text{min}} \leq \|w_i\|_2 \leq \delta_{\text{max}}\}$, which better accounts for the heavy tails observed in the latent space distribution. The optimization problem for PULSE++ can be formalized as:

$$\hat{w}, \hat{n} = \arg\min_{w,n} \|I - G(w,n)\|_2^2 + R(w,n), \quad \text{s.t. } w_i \in A, \forall i \in \{1,\ldots,l\}, \tag{3.33}$$

where $I \in \mathbb{R}^m$ represents the input projections, $G(w,n) \in \mathbb{R}^m$ is the generative model output, and $R(w,n)$ is the regularization term, given by:

$$R(w,n) = \lambda \text{Cross}(w) + \frac{1}{2}\sum_{i=1}^{l} \|n_i\|_2^2, \tag{3.34}$$

where $w = \{w_1, w_2, \ldots, w_l\} \in \mathbb{R}^{l \times k}$ are the latent vectors, $n = \{n_1, n_2, \ldots, n_l\} \in \mathbb{R}^l$ are the noise vectors, $\text{Cross}(w)$ represents the pairwise Euclidean distance between latent vectors w_i and w_j, λ is a hyperparameter controlling the trade-off between data consistency and maintaining the structure of the latent space.

Numerical experiments were conducted to evaluate the performance of PULSE++ on two tomographic imaging systems: one using incomplete Fourier space measurements for MRI and another based on a limited-angle X-ray fan-beam CT system. For the CT experiments, 2D CT slices (512×512 pixels) were simulated using projection data from 120 views. Noiseless X-ray measurements were modeled using the Beer-Lambert law, with additional controlled noise following a Poisson distribution.

Similar to PULSE [Menon, 2020] and BRGM [Marinescu, 2020], PULSE++ can find multiple solutions that satisfy the measurements. Using the Adam optimizer, PULSE++ explores different possible solutions through empirical sampling, as in [Menon, 2020], by restarting the optimization multiple times. This iterative process continues until it finds a solution $G(\hat{w}, \hat{n})$ that meets the data fidelity condition within a tolerance level ϵ_n, adjusted for the measurement noise n. The alternate solutions found by empirical sampling show

significant variability in fine structures, demonstrating the method's capacity to capture multiple plausible reconstructions for high-dimensional objects.

Additionally, uncertainty maps, computed as the pixel-wise standard deviation across alternate solutions, reveal areas of uncertainty in the reconstructions. These maps indicate that PULSE++ enforces strong data consistency, shown by lower uncertainty in key regions of the reconstructed images.

PULSE++ has proven robust in producing data-consistent solutions, though its accuracy remains influenced by the quality of the StyleGAN model used to represent the object distribution. In medical imaging, this can introduce representation errors if certain pathologies are underrepresented or if the training data is biased toward specific populations.

This work represents the first application of compressed sensing with generative models for sparse 2D tomographic reconstruction, yielding promising results. However, its application has yet to extend to more complex tasks, such as 3D imaging with very limited projections or challenging geometries like 3D reconstruction from partial biplanar cone-beam X-rays. Our research aims to tackle these challenges.

3.3 3D Reconstruction from Biplanar X-Rays

With very few projections, it is very difficult to disentangle the structures for even coarse 3D estimation. In other words, many 3D volumes may have generated such projections *a priori*. As discussed in the previous section, classical analytical and iterative methods struggle to provide even close reconstructions when the number of available projections is very limited.

Several deep learning approaches have focused on significantly reducing the number of X-ray projections required for accurate volumetric reconstruction. Figure 3.17 illustrates the different paradigms for this task. Initially, as with other inverse problems in computer vision, the focus was on feedforward methods adapted for tomographic reconstruction.

At the time of developing our method, the predominant approaches were these feedforward models such as those proposed by [Henzler, 2018; Shen, 2019; Ying, 2019; Shen, 2022b; Jiang, 2022]. These models aimed to invert projections to predict 3D volumes directly from a minimal number of projections—sometimes using only one or two. Most of these methods used feature embeddings from 2D projections [Henzler, 2018; Shen, 2019], often combined with fusion mechanisms [Ying, 2019; Jiang, 2022; Tan, 2022; Lu, 2022; Zhang, 2023b] and adversarial losses [Ying, 2019; Jiang, 2022; Wang, 2023]. Some approaches further constrained the solution space with geometric constraints and 3D refinement networks, such as U-Net [Shen, 2022b; Lu, 2022].

Henzler et al. [Henzler, 2018] were the first to drastically reduce the number of projections needed for 3D reconstruction to just one, focusing on 3D cranial bone estimation. They demonstrated that deep learning-based CNNs could handle this challenging task by using data-driven priors. Their method involved training on a large paired dataset of syn-

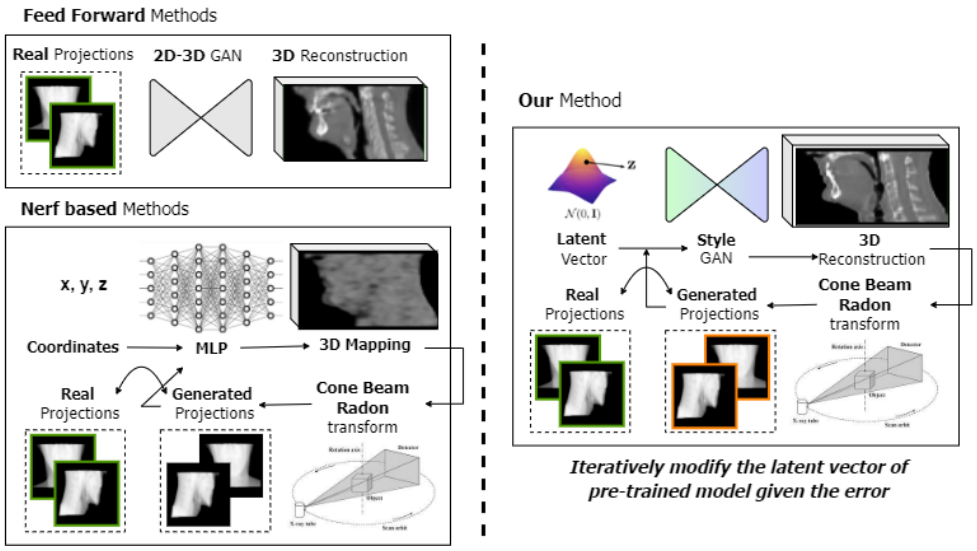

Figure 3.17: **Current methods vs our method.** Feed-forward methods do not manage to predict a detailed and matching tomographic volume from a few projections. Iterative methods based on neural radiance fields lack prior for good reconstruction. By learning an embedding for the possible volumes, we can recover an accurate volume from very few projections with an optimization based on a Bayesian formulation.

thetic projections and 3D volumes to learn how to invert a 2D projection into 3D. First, the model generated a coarse, low-resolution volume, which was then refined by fusing it with the input X-ray to create a high-resolution 3D volume. This approach achieved results that would have been nearly impossible for traditional, non-deep learning methods with only a single X-ray.

Building on this idea, Shen et al. [Shen, 2019] extended the method to CT generation. They also relied on large paired datasets of synthetic projections and CT volumes, using again data-driven priors to reconstruct 3D volumes from a single projection. Their model used 2D feature embedding and 3D decoding to map projection radiographs to corresponding 3D anatomy, showing its effectiveness with CT scans of the upper abdomen, lung, and head-and-neck regions. This approach demonstrates the potential for creating volumetric tomographic X-ray images from just one projection.

While using just one projection is promising, it is still far from producing usable CT-quality reconstructions. A significant amount of ambiguity remains, with multiple possible solutions for the 3D structure. X2CT [Ying, 2019] extended this approach by using biplanar X-rays, where orthogonal projections provide more anatomical constraints by resolving some depth ambiguities through a lateral view. In this method, as we can see in Figure 3.18, the conditional GAN framework is used to reconstruct CT volumes from two orthogonal X-rays. A designed generator network increases the dimensionality from 2D to 3D. A novel feature fusion method is introduced to combine the information from both X-rays. The model is trained using a combination of mean squared error (MSE) loss and adversarial loss with projection loss, resulting in better and more realistic 3D CT

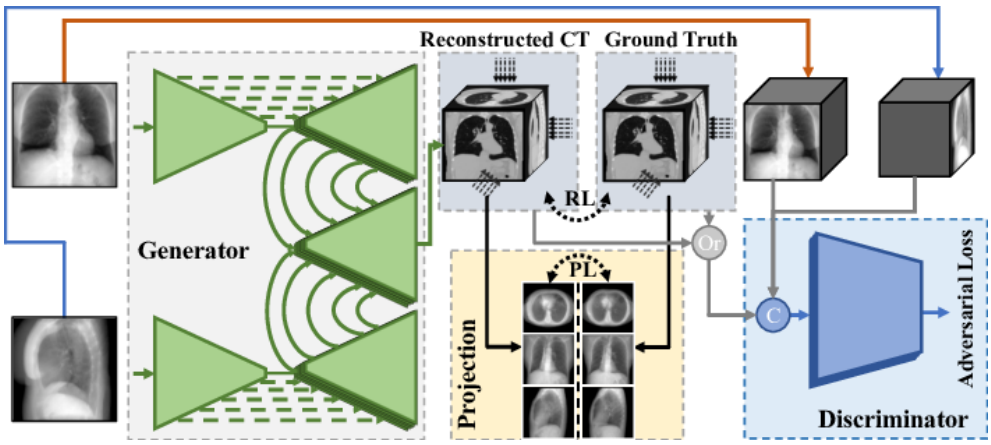

Figure 3.18: X2CT-GAN [Ying, 2019] framework. The model encodes features from biplanar X-rays, fuses them, and decodes them into 3D volumes. The training process is guided by a combination of MSE, an adversarial loss, and a projection loss, to ensure better realism and geometric coherence with the target.

reconstructions. Experiments were done on chest CT dataset and showed effectiveness of the method compared to previous work.

Shen's second approach [Shen, 2022b] revisits classical methods by introducing geometric constraints, recognizing that the 2D-to-3D feature transformation lacks inherent geometric consistency. To address this, the method leverages classical backprojection to enforce 3D geometric constraints, even with very few projections. First, it attempts to predict missing projections by disentangling content and style from the available projections. After generating these additional views, both the real and predicted projections are used in backprojection. Instead of using traditional filtering methods like in FBP, the model employs a 3D U-Net to remove artifacts and predict the 3D internal structures. This method combines classical and deep learning approaches to tackle this extreme sampling for better results. Yet, high quality and accurate alignment is to be reached.

Recently, more advanced architectures have been used for better 2D feature embedding, using self-attention mechanisms [Tan, 2022] and transformer-based networks [Zhang, 2023b; Wang, 2023]. Tan et al. [Tan, 2023] also explored semi-supervised learning using a teacher-student framework to address the challenge of limited paired volume-projection data.

However, these feedforward approaches can struggle with previously mentioned averaging effects [Menon, 2020], where all possible solutions are averaged, leading to blurred and unmatching reconstructions. Additionally, they may not ensure data fidelity by being consistent with the original projections, and they can generalize poorly.

Other methods have adapted NeRFs [Mildenhall, 2021] for tomographic reconstruction with sparsely sampled data, typically using 20 to 30 projections, with fan-beam or cone-beam 2D or 3D reconstruction[Zha, 2022; Shen, 2022a]. Approaches like [Zha, 2022; Shen, 2022a] iteratively optimize the reconstructed volume based on available projections,

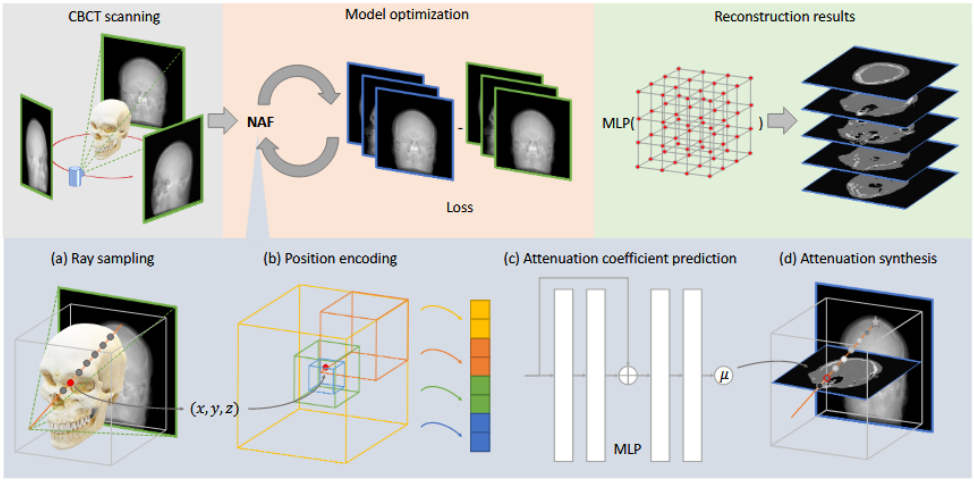

Figure 3.19: Nerf-based NAF pipeline [Zha, 2022]. The gray block shows how the CBCT scanner captures X-ray projections from different views. The blue block simulates projections using NAF, while the orange block represents the optimization process where NAF is refined by comparing real and synthesized projections. Finally, the green block illustrates how NAF generates a CT model by querying corresponding voxels. This iterative pipeline performs well for sparse projections but lacks sufficient priors when very few projections are available.

generally achieving better results by ensuring projection consistency and improving generalization. Building on the deep image prior [Ulyanov, 2018] of MLPs, they can reduce the need for measurements without relying on sparsity or smoothness, as shown by [Shen, 2022a]. However, these methods without prior resemble classical iterative reconstruction techniques without strong regularization and can fail if insufficient constraints are provided. Additionally, [Shen, 2022a] can initialize NeRF using a pre-acquired volume, but this approach tends to fail when very few projections are available, as our experiments will demonstrate, as shown in Section 3.5.3. Recently, gaussian splatting [Kerbl, 2023] methods have also been developed following the same idea with improved results [Lin, 2024].

To surpass previous feedforward and iterative methods that lacked priors, we built on the best of both worlds. We introduced the first method for 3D reconstruction from very few X-rays that builds on the legacy of compressed sensing with deep generative models, using posterior sampling to find the optimal solution by relying directly on the manifold of possible solutions. We named our method X2Vision.

As illustrated in Figure 3.17, to be able to reconstruct a volume accurately given as low as two projections only, we first need to learn a prior on the volume. To address this ill-posed problem, we introduce prior knowledge of anatomic structures by training a generative model on 3D CTs of head and neck. To do this, we leverage the potential of generative models to learn a low-dimensional manifold of the 3D target body part.

We optimize the latent vectors of the generative model to recover a volume that both integrates this prior knowledge and ensures consistency between the reconstructed

image and input projections. Given projections, we find by the Bayesian formulation, the intermediate latent vectors conditioning the generative model that minimize the error between synthesized projections of our reconstruction and these input projections.

2D Experiment We initially focused on 2D by training a StyleGAN2 [Karras, 2020b] model on individual 2D slices of head and neck CT scans to learn the anatomical manifold. By conditioning the model on the cranio-caudal slice position, we enabled it to learn localized manifolds specific to each region. Once trained, the model was used to reconstruct each slice independently from only two fan-beam biplanar projections, effectively capturing diverse and realistic anatomical structures within each slice.

However, reconstructing slices independently led to inconsistencies across slices. The Appendix 3.7 includes images illustrating both the strengths and limitations of this 2D approach, highlighting the realistic structures generated as well as the inconsistencies between slices.

Maintaining consistency across slices is challenging. Some methods address this by applying a final cross-sectional correction step or developing a 3D prior, while others simultaneously update all slices to enhance coherence. We explored these alternatives and found that directly using a cross-sectional prior provided the most effective solution, especially for partial cone-beam projections.

Encouraged by the potential of the 2D approach, we advanced to 3D modeling to achieve improved results through a global prior, unified optimization, and more consistent reconstructions.

To extend our work into 3D, we build on the state-of-the-art generative model StyleGAN2 [Karras, 2019; Karras, 2020b], adapted for 3D by Hong et al. [Hong, 2021], which we further enhance with a more complex network and training framework. Compared to other 3D GANs, this model has demonstrated superior disentanglement of semantic features in the feature space [Ellis, 2022]. Figure 3.20 presents the architecture of the general 3D StyleGAN [Hong, 2021].

By relying directly on the manifold, our method avoids averaging effects, produces more accurate and realistic reconstructions, and ensures better alignment and consistency with the input projections. In contrast to feedforward methods, which are tied to a specific projection geometry calibration and require paired projection-reconstruction data, our approach can be used with varying numbers of projections and different projection geometries without the need for retraining. Compared to NeRF-based methods, our method exploits prior knowledge from many patients to require only two projections.

We evaluate our method on reconstructing cancer patients' head-and-neck CTs, which involves intricate and complicated structures, as presented in Chapter 1. We perform several experiments to compare our method with a feedforward-based method [Ying, 2019] and a recent NeRF-based method [Shen, 2022a], which are the previous state-of-the-art methods for the very few or few projections cases, respectively.

We show that our method allows to retrieve results with the finest reconstructions

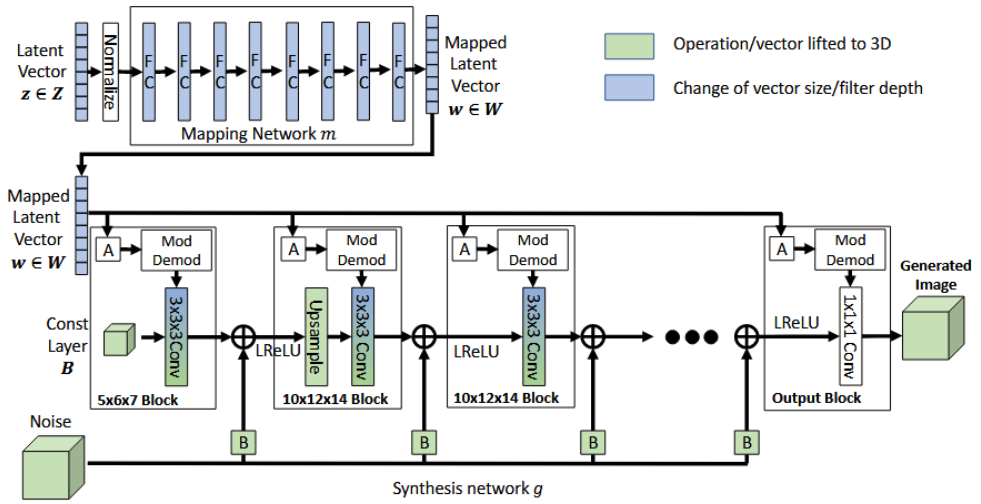

Figure 3.20: Architecture of 3D StyleGAN [Hong, 2021]. Adapted from the StyleGAN architecture by extending 2D components to 3D.

and better matching structures, for a variety of number of projections. To summarize, our contributions are two-fold: (i) A new paradigm for 3D reconstruction with biplanar X-rays: instead of learning to invert the measurements, we leverage a 3D style-based generative model to learn deep image priors of anatomic structures and optimize over the latent space to match the input projections; (ii) A novel unsupervised method, fast and robust to sampling ratio, source energy, angles and geometry of projections, all of which making it general for downstream applications and imaging systems.

3.4 X2Vision

Figure 3.21 gives an overview of the pipeline we propose. We first learn the low-dimensional manifold of CT volumes of a target body region. At inference, we estimate the MAP volume on this manifold given very few projections: we find the latent vectors that minimize the error between the synthetic projections from the corresponding volume on the manifold and the real ones. In this section, we formalize the problem, describe how we learn the manifold, and detail how we optimize the latent vectors.

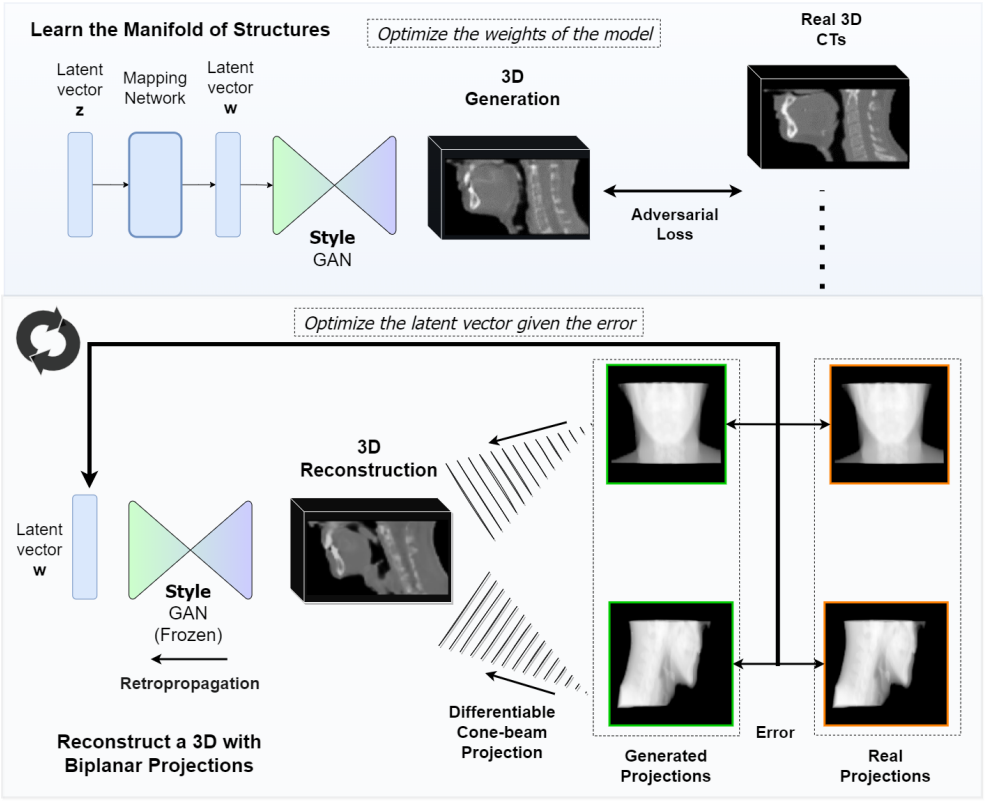

Figure 3.21: **Our pipeline.** We first learn the low-dimensional manifold of 3D structures using a generative model. Then, given projections, we find the latent vectors that minimize the error between the projections of our generation and the input projections.

3.4.1 Problem Formulation

Given a small set of projections $\{I_i\}_i$, possibly as few as two, we would like to reconstruct the 3D tomographic volume v that generates these projections. This is a hard ill-posed problem, and to solve it, we need prior knowledge about the possible volumes. To do this, we look for the MAP estimate given the projections $\{I_i\}_i$:

$$v^* = \underset{v}{\operatorname{argmax}}\, p(v|\{I_i\}_i) = \underset{v}{\operatorname{argmax}}\, p(v)p(\{I_i\}_i|v) = \underset{v}{\operatorname{argmin}} \sum_i \mathcal{L}(v, I_i) + R(v)\,. \quad (3.35)$$

Term $\mathcal{L}(v, I_i)$ is a log-likelihood. We take it as:

$$\mathcal{L}(v, I_i) = \lambda_2 \left\| A_i \circ v - I_i \right\|_2 + \lambda_p \mathcal{L}_p(A_i \circ v, I_i)\,, \quad (3.36)$$

where A_i is an operator that projects volume v under view i. We provide more details about operator A in Section 3.4.3. $\mathcal{L}_p$ is the perceptual loss [Johnson, 2016] between projection of v and the observed projection I_i. Term $R(v)$ is a regularization term. It is crucial as it is the term that embodies prior knowledge about the volume to reconstruct.

As discussed in the introduction, we rely on a generative model, which we describe in the next section. Then, we describe how exactly we use this generative model for regularization term $R(v)$ and how this changes our optimization problem.

3.4.2 Manifold Learning

To regularize the domain space of solutions, we leverage a style-based generative model to learn deep priors of anatomic structures and approximate the domain space of CTs. Our model relies on StyleGAN2 [Karras, 2020b] that we extend in 3D by changing the 2D convolutions into 3D ones as done in 3DStyleGAN [Hong, 2021] except that we start from the StyleGAN2 architecture.

Our generator G generates a volume v given a latent vector $\mathbf{w}$ and Gaussian noise vectors $\mathbf{n} = \{\mathbf{n}_j\}_j$: $v = G(\mathbf{w}, \mathbf{n})$. Latent vector $\mathbf{w} \in \mathcal{N}(\mathbf{w}|\mu, \sigma)$ is computed from an initial latent vector $\mathbf{z} \in \mathcal{N}(0, I)$ mapped using a learned network m: $\mathbf{w} = m(\mathbf{z})$. $\mathbf{w}$ controls the global structure of the predicted volumes at different scales by its components $\mathbf{w}_i$, while the noise vectors $\mathbf{n}$ allow more fine-grained details. The mean μ and standard deviation σ of the mapped latent space can be computed by mapping over initial latent space $\mathcal{N}(0, I)$ after training. The mapping network learns to disentangle the initial latent space relatively to semantic features which is crucial for the inverse problem. We train this model using the non-saturating logistic loss [Goodfellow, 2020] and path length regularization [Karras, 2020b]. For the discriminator, we use the non-saturating logistic loss with R1 regularization [Mescheder, 2018]. We implement adaptive discriminator augmentation from StyleGAN-ADA [Karras, 2020a] to improve learning of the model's manifold with limited medical imaging data.

3.4.3 Reconstruction from Biplanar Projections

Since our generative model provides a volume v as a function of vectors $\mathbf{w}$ and $\mathbf{n}$, we can reparameterize our optimization from Eq. (3.35) into:

$$\mathbf{w}^*, \mathbf{n}^* = \underset{\mathbf{w},\mathbf{n}}{\operatorname{argmin}} \sum_i \mathcal{L}(G(\mathbf{w}, \mathbf{n}), I_i) + R(\mathbf{w}, \mathbf{n}) \,. \tag{3.37}$$

This formalism builds on the one introduced by [Marinescu, 2020]. Similarly, we optimize the latent vectors $[w_1, w_i, \dots]$ independently in the $\mathcal{W}^+$ space. Note that by contrast with [Marinescu, 2020], we optimize on the noise vectors $\mathbf{n}$ as well: as we discovered in our early experiments, the $\mathbf{n}$ are also useful to embed high-resolution details.

We take our regularization term $R(\mathbf{w}, \mathbf{n})$ as :

$$R(\mathbf{w}, \mathbf{n}) = \lambda_w \mathcal{L}_w(\mathbf{w}) + \lambda_c \mathcal{L}_c(\mathbf{w}) + \lambda_n \mathcal{L}_n(\mathbf{n}) \,. \tag{3.38}$$

Term $\mathcal{L}_w(\mathbf{w}) = -\sum_k \log \mathcal{N}(\mathbf{w}_k|\mu, \sigma)$ ensures that $\mathbf{w}$ lies on the same distribution as during training. $\mathcal{N}(\cdot|\mu, \sigma)$ represents the density of the standard normal distribution of

mean μ and standard deviation σ.

Term $\mathcal{L}_c(\mathbf{w}) = -\sum_{i,j} \log \mathcal{M}(\theta_{i,j}|0,\kappa)$ encourages the $\mathbf{w}_i$ vectors to be collinear so to keep the generation of coarse-to-fine structures coherent. $\mathcal{M}(\cdot;\mu,\kappa)$ is the density of the Von Mises distribution of mean μ and scale κ, which we take fixed, and $\theta_{i,j} = \arccos(\frac{\mathbf{w}_i \cdot \mathbf{w}_j}{\|\mathbf{w}_i\|\|\mathbf{w}_j\|})$ is the angle between vectors $\mathbf{w}_i$ and $\mathbf{w}_j$.

Term $\mathcal{L}_n(\mathbf{n}) = -\sum_j \log \mathcal{N}(\mathbf{n}_j|\mathbf{0}, I)$ ensures that the $\mathbf{n}_j$ lie on the same distribution as during training, *i.e.*, a multivariate standard normal distribution. The λ_* are fixed weights.

Finally, we obtain the MAP estimate for the 3D reconstruction from biplanar X-rays as:

$$v^* = G(w^*, n^*). \tag{3.39}$$

Projection Operator.

In practice, we take operator A as a 3D cone beam projection that simulates X-ray attenuation across the patient, called DRR, adapted from DeepDRR [Unberath, 2018] and XraySyn[Peng, 2021]. We model a realistic X-ray attenuation as a ray tracing projection using material and spectrum awareness, derived from 3.2:

$$\mathcal{I}_{\text{atten}} = \sum_E \mathcal{I}_0 e^{-\sum_m \mu(m,E) t_m}, \tag{3.40}$$

with $\mu(m,E)$ the linear attenuation coefficient of material m at energy state E that is known [Hubbell, 1995], t_m the material thickness, $\mathcal{I}_0$ the intensity of the source X-ray. For materials, we consider the bones and tissues that we separate by threshold on electron density. At this stage, we consider only the primary ray and omit additional noise or scattering effects, as the problem is already highly ill-posed. A inverts the attenuation intensities $\mathcal{I}_{\text{atten}}$ to generate an X-ray along few directions successively. We make A differentiable using [Peng, 2021] to allow end-to-end iterative optimization for reconstruction.

The projector utilizes Siddon's algorithm [Siddon, 1985], optimized with CUDA on a GPU, to efficiently trace the X-ray paths through the volume and compute the corresponding attenuations.

3.5 Experiments and Results

3.5.1 Dataset and Preprocessing

Manifold Learning. We trained our model with a large dataset of 3500 CTs of patients with head-and-neck cancer (including contrast-enhanced and non contrast-enhanced), more exactly 2297 patients from the publicly available The Cancer Imaging Archive (TCIA) [Grossberg, 2020; Kwan, 2019; Vallières, 2020; Beichel, 2015; Kinahan, 2020; Zuley, 2015] and 1203 from private internal data, after obtention of ethical approbations. We split this data into 3000 cases for training, 250 for validation, and 250 for testing. We focused CT scans on the head and neck region above shoulders, with a resolution of $80 \times 96 \times 112$ ($1.3 \times 2.4 \times 1.9mm^3$), and centered on the mouth after automatic segmentation using a pre-trained U-Net [Ronneberger, 2015]. This allowed us to concentrate on the central zone of the head and neck to demonstrate the feasibility of 3D learning and reconstruction with biplanar X-rays. The CTs were preprocessed by min-max normalization after clipping between -1024 and 2000 HUs.

3D Reconstruction. With patient consent, we compiled planning CT scans (contrast-enhanced) and subsequent CBCT scans from 242 patients across two medical centers (CLB and IGR), one contributing 177 and the other 65 cases. These datasets are distinct in protocols and scanning equipment. All CBCTs were acquired on imager of VersaHD from Elekta [Elekta, 2023]. To evaluate our approach, we sampled 80 patients from the first cohort.

As depicted in Figure 3.22, notable differences emerge between the initial CT scans and subsequent CBCT scans, because of both treatment-induced alterations and patient pose variations. To compare our reconstruction in the calibrated HU space, we registered the planning CTs on the CBCTs by deformable registration with MRF minimization [Glocker, 2008]. We hence obtained 3D volumes as virtual CTs we considered as ground truths for our reconstructions after normalization. From these volumes, we generated projections using the projection module described in Section 3.4.3.

3.5.2 Implementation Details

Manifold Learning. We used Pytorch [Paszke, 2019] to implement our model, based on StyleGAN2 [Karras, 2020b] extended in 3D. It has a starting base layer of $256 \times 5 \times 6 \times 7$ and includes four upsamplings with 3D convolutions and filter maps of $256, 128, 64, 32$. We also used 8 fully-convolutional layers with dimension 512 and an input latent vector of dimension 512, with tanh function as output activation. To optimize our model, we used lazy regularization [Karras, 2020b] and style mixing [Karras, 2020b], and added a 0.2 probability for generating images without Gaussian noise to focus on embedding the most information. We augmented the discriminator with vertical and depth-oriented flips, rotation, scaling, motion blur and Gaussian noise at a probability of 0.2. Our training used

mixed precision on a single GPU Nvidia Geforce GTX 3090 with a batch size of 6, and we optimized the generator, discriminator, and mapping networks using Adam at learning rates $6e{-5}$ and $1e{-5}$ to avoid mode collapse and unstable training. After training for 4 weeks, we achieved stabilization of the Fréchet Inception Distance (FID) [Heusel, 2017] and Multi-scale Structural Similarity (MS-SSIM) [Wang, 2003] on the training set.

3D Reconstruction. For the reconstruction, we performed the optimization on GPU V100 PCI-E using gradient descent with Adam, with learning rate of $1e{-3}$. By grid search on the validation set, we selected the best weights that well balance between structure and fine-grained details, $\lambda_2 = 10$, $\lambda_p = 0.1$, $\lambda_w = 0.1$, $\lambda_c = 0.05$, $\lambda_n = 10$.

We compute μ and σ by taking the mean and standard deviation of $10,000$ latent variables passed through the mapping network, similar to the method used in the original StyleGAN2 inversion [Karras, 2020b]. We perform 100 optimization steps starting from the mean of the mapped latent space μ, which takes 25 seconds, enabling potential clinical use like pre-treatment positioning or adaptation of treatment.

3.5.3 Results and Discussion

Manifold Learning. We tested our model's ability to learn the low-dimensional manifold. We used FID [Heusel, 2017] to measure the distance between the distribution of generated volumes and real volumes, and MS-SSIM [Wang, 2003] to evaluate volumes' diversity and quality. We obtained a 3D FID of 46 and a MS-SSIM of 0.92. For reference, compared to 3DStyleGAN [Hong, 2021], our model achieved half their FID score on another brain MRI dataset, with comparable MS-SSIM. This may be due to a more complex architecture, discriminator augmentation, or simpler anatomy.

We further evaluated our model's ability to generate and retrieve realistic 3D CT scans of the head and neck region. After training, we randomly selected latent vectors to generate diverse and realistic 3D CTs, as shown in the Appendix 3.7. Additionally, when given an unseen CT, the model was able to generate the closest artificial version by projecting it onto the learned manifold, producing a synthetic 3D CT that closely resembles the real one. Our model demonstrates a good capacity for capturing the diversity of anatomies with matching details. On average, we achieved a representation error [Bora, 2017] of 1.7% (std. 0.5%) (normalized MAE) across 60 test patients.

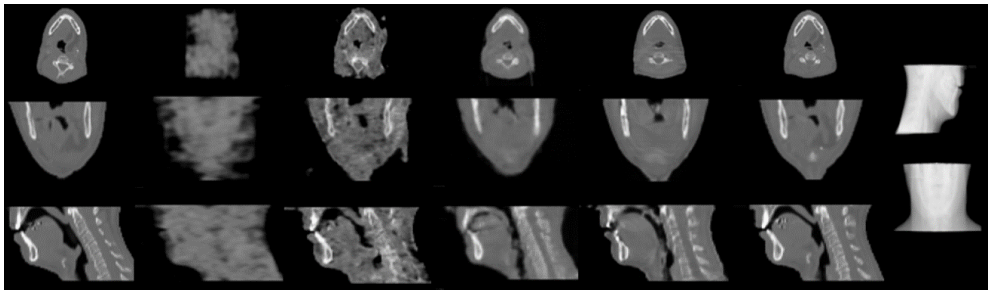

Figure 3.22: **Visual comparison of 3D reconstruction from biplanar projections by our model and baselines.** Without a previous CT volume, NeRP fails by lack of constraints. When initialized with an earlier CT (left), NeRP tends to create artifacts to match the projections rather than really change the anatomy. Our method produces better matching and less blurred structures than X2CT-GAN, almost matching the CT volume deformed on the CBCT volume (GT, right).

Baselines. We compared our method against the main feed-forward method X2CT-GAN [Ying, 2019] and the neural radiance fields with prior image embedding method NeRP [Shen, 2022a] meant for modest sparsely-sampled reconstruction. Recent methods like [Shen, 2022b] and [Jiang, 2022] were excluded because they provide only minor improvements compared to X2CT-GAN [Ying, 2019] and have similar constraints to feed-forward methods. Additionally, no public implementation is available.

3D Reconstruction. Figure 3.22 compares our reconstruction with those of the baselines from biplanar projections. Appendix shows further visual comparisons. Our approach better fits the patient's anatomical structures, including bones, tissues, and air separations, closely matching the real CT volume and providing more realistic reconstructions. In contrast, X2CT-GAN [Ying, 2019] produced generally realistic structures but with blurriness and failed to align accurately with actual structures, as it does not enforce consistency with the projections. The blurring is an effect of averaging multiple possible solutions, whereas our method selects a solution on the learned manifold, resulting in a more realistic and detailed reconstruction.

In some clinical procedures, an earlier CT volume of the patient may be available and can be used as an additional input for NeRP [Shen, 2022a]. Without a previous CT volume, NeRP lacks the necessary prior to accurately solve the ill-posed problem. Even when initialised with a previous CT volume, NeRP often fails to converge to the correct volume and introduces many artifacts when very few projections are used. In contrast, our method introduces the required structure prior and produces better results.

We used two quantitative metrics—Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM)—to assess reconstruction error and perceptual quality, respectively. Table 3.1 shows these metrics for our method and baselines with 1 to 8 cone beam projections. Deviation from projections, as in X2CT-GAN, leads to inaccurate

Table 3.1: **Metrics for our method and baselines, for reconstruction from 1 to 8 cone beam projections.** Standard deviations are provided in parentheses.

Method	1 Projection		**2 Projections**	
	PSNR (dB)↑	SSIM↑	**PSNR (dB)↑**	**SSIM↑**
NeRP (w/o prior volume)	14.8 (±2.7)	0.12 (±0.10)	18.4 (±3.8)	0.17 (±0.10)
NeRP (w/ prior volume)	22.5 (±3.2)	0.29 (±0.07)	23.5 (±3.5)	0.30 (±0.06)
X2CT-GAN	20.7 (±2.4)	0.57 (±0.07)	21.8 (±2.5)	0.72 (±0.08)
Ours	**23.2** (±**2.8**)	**0.79** (±**0.09**)	**25.8** (±**3.2**)	**0.85** (±**0.10**)
	4 Projections		8 Projections	
NeRP (w/o prior volume)	19.9 (±2.6)	0.21 (±0.04)	20.0 (±2.5)	0.23 (±0.05)
NeRP (w/ prior volume)	24.2 (±2.7)	0.32 (±0.05)	24.9 (±4.9)	0.34 (±0.08)
Ours	**28.2** (±**3.5**)	**0.89** (±**0.10**)	**30.1** (±**3.9**)	**0.92** (±**0.11**)

reconstruction. However, relying solely on projection consistency is inadequate for this ill-posed problem. NeRP matches projections but cannot reconstruct the volume correctly. Our approach balances between instant and iterative methods by providing a reconstruction in 25 seconds with 100 optimization steps, while ensuring maximal consistency. In contrast, NeRP requires 7 minutes, and X2CT-GAN produces structures instantly but unmatching. Clinical CBCT acquisition and reconstruction by FDK [Feldkamp, 1984a] take more than 2 minutes and 10 seconds respectively. Our approach significantly reduces clinical time and radiation dose by using instant biplanar projections, making it promising for fast 3D visualization towards enabling complex positioning and adaptive planning.

Clinical Evaluation

To provide a more clinically relevant evaluation, we assessed the performance of our model using clinical metrics. Specifically, we evaluated the model's 3D rigid registration accuracy by comparing the six degrees of freedom (translation and rotation) parameters obtained when registering the planning CT to our reconstructions with those from the ground truth (CT deformed to match CBCT). On average, we achieved translation errors of 0.45 mm (±0.31 mm) and rotation errors of 0.50° (±0.26°) across all axes. This demonstrates that our reconstructions can be accurately registered, showing performance similar to 3D CBCT or ground truth data.

We further tested our model in a more clinical setting by generating full field-of-view CBCT projections that resemble real CBCT projections. Since our training was conducted on partial field-of-view data with central fixed positioning—while CBCT is centered on the PTV—we adapted our reconstruction process accordingly. We first performed pre-positioning to set the isocenter, applied projection masking for regions of interest, and then optimized the reconstruction.

Visual results (Figure 3.23) show that our reconstruction closely aligns with the coarse structure of the paired CBCT. Despite using 100 times fewer projections, our reconstruction achieves nearly CBCT-level quality (bottom left). The anatomical accuracy of the

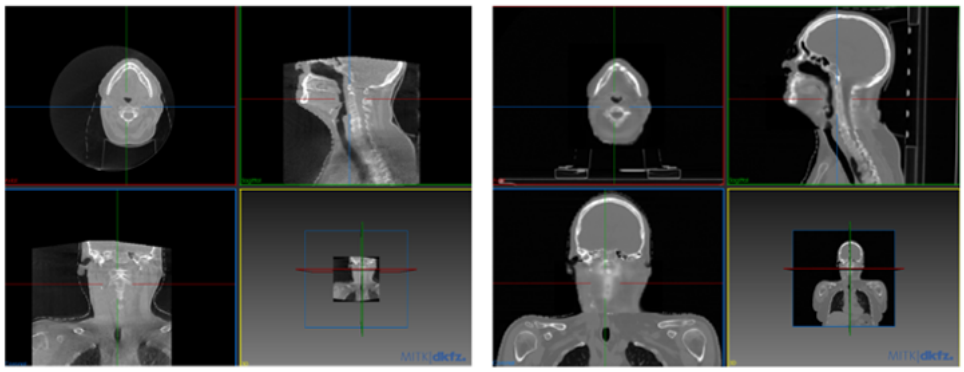

Figure 3.23: Example of 3D CT reconstruction using two orthogonal cone-beam projections (in-field), extended with rigid registration of out-of-field regions from the planning CT (right), compared to the original CBCT (left). The reconstruction achieves good anatomical fidelity with significantly fewer projections, enabling accurate planning CT registration and completion of out-of-field areas (right).

3D reconstruction allows for precise registration of the planning CT, enabling completion of the reconstruction for out-of-field regions (bottom right). This could facilitate full dosimetry by providing sufficient margin views of surrounding structures and the presence of shoulders to simulate beams targeting lymph nodes—potentially enabling comprehensive dosimetric evaluations.

We also compared the translation and rotation parameters obtained from registering the planning CT with those derived from our reconstruction and those obtained clinically via CBCT. On average, we observed deviations of 0.4 mm in translation and 0.2° in rotation across all axes. These results demonstrate the potential to enhance biplanar X-rays by providing 3D fine tissue registration comparable to that of CBCT, potentially reducing the need for CBCT in positioning.

Additionally, to demonstrate the potential for dose accumulation and treatment adaptation based on our reconstruction, we conducted a preliminary dosimetry evaluation on three test cases using the corresponding initial treatment plans. VMAT was simulated with an in-house method using the collapsed cone algorithm [Ahnesjö, 1989], utilizing their respective plans and projected target and OARs volumes from the ground truth. The gamma index and dose-volume histogram (DVH) comparisons for several organs showed favorable results, with an average gamma index of 98% at 3mm/3% and 96% at 3mm/2%.

Further details are provided in the Appendix: Figure 3.30 presents an example of the reconstruction, Figure 3.31 displays the corresponding simulated dose maps, Figure 3.32 shows the gamma index map at 3 mm, and Figure 3.33 illustrates the DVH comparison based on these dose distributions. These preliminary results indicate that the reconstructed structures align well in terms of density and anatomical accuracy, suggesting the potential for dose accumulation and triggering the need for adaptation based on our reconstructions.

3.6 Conclusion and Discussion

We have proposed an unsupervised method for 3D reconstruction from biplanar X-rays using a deep generative model. By learning the structural manifold and retrieving the maximum a posteriori volume from the projections, our approach achieves state-of-the-art reconstruction performance. This method is fast, robust, and adaptable across different anatomical regions, making it suitable for various clinical applications, including patient positioning and visualization with reduced radiation exposure.

Our approach can reconstruct a full CT from only partial biplanar images, achieving quality comparable to CBCT while preserving good tissue attenuation. This capability accelerates on-board patient positioning and has the potential to support adaptive radiotherapy, offering a significant clinical advantage by providing a fast, low-dose, and cost-effective alternative to CBCT.

Clinical Translation Our method has the potential to replace CBCT for guiding adaptive radiotherapy by enabling 3D reconstructions of detailed structures and tissues. This allows for more accurate positioning than 2D X-rays alone and supports adaptive treatment adjustments.

For successful clinical translation, thorough evaluation of clinical metrics and outcomes is essential, particularly through comprehensive dosimetry validation for 3D adaptive radiotherapy. Preliminary results indicate that our method accurately captures the key structures necessary for dosimetry. Ensuring reliable dose calculations and accumulation based on these reconstructions could facilitate timely adaptations in treatment plans according to the reconstructed anatomy.

While our method demonstrates a high level of alignment with ground truth data, further improvements in quality and robustness are necessary for clinical application. Certain detailed structures—such as cervical details, muscle and internal organ distribution, and tumor contours— may not fully be captured by the current model. Additionally, these structures may be present but challenging to reconstruct accurately from just two projections, as minor shifts in 2D projections can lead to significant variations in global structure and result in the loss of fine density details.

Accurate tumor reconstruction is crucial for adaptive replanning, where precise delineation is necessary. Although our model has been tested on patients with tumors, it has not yet been fully validated specifically for tumor reconstructions. Tumors are particularly difficult to identify on non-contrast CT and even more so to reconstruct from non-contrast X-rays used in radiotherapy, where even CBCT, with significantly more projections, often struggles with tumor visibility. To improve accuracy, a comprehensive dataset of contrast-enhanced CTs is needed. In this study, we trained our model on both contrast-enhanced and non-contrast CTs to assess feasibility. However, it is important to note that tumor reconstruction from just two projections may rely more on learned statistical patterns than on precise anatomical recovery, given the complexity and inherent ambiguity of the task.

Achieving perfect segmentation of OARs and tumors based solely on our reconstructions may not be feasible, limiting the potential for direct replanning using these images. However, achieving clinical outcomes such as accurate contouring and dosimetry simulations for replanning may benefit from applying deformable registration techniques from the planning CT to propagate structures like OARs and tumors, similar to current CBCT-based methods, presented in Chapter2.

To assess the feasibility of accurate 3D reconstruction from biplanar X-rays and address the challenges of this highly ill-posed problem, we generated projections that perfectly align with 3D structures. However, extending this approach to real clinical settings, with considerations for calibration, scatter effects, noise, and patient positioning variability, is essential. A preliminary method for this extension will be presented in the final chapter 5. We propose validating our approach with real biplanar X-rays from the ExacTrac biplanar system, using actual projections paired with corresponding CBCT data for clinical evaluation.

Finally, while rigid registration can help extend to out-of-field regions, its effectiveness is limited by non-rigid anatomical changes, which create uneven transitions with local reconstructions and inconsistencies in patient anatomy outside the field. Instead, using an expanded generative model trained on full head and neck anatomy could provide a more seamless and realistic extension from the given projections. This approach will be further discussed in the final chapter 5.

Unsupervised Approach. A significant advantage of our method is that it does not require training with paired projection-3D data. Instead, we leverage state-of-the-art 3D generative models and invert them for 3D reconstruction. Pre-trained models can be adapted for many downstream tasks and used solely during the test phase. This approach allows us to explore and utilize better pre-trained models, with the potential to use a single model for large anatomical regions—such as the head and neck—to solve various inverse problems either independently or simultaneously.

Our method is also independent of the number of X-ray projections, calibration settings, and machine geometry. Therefore, it does not require retraining if X-ray machine settings, such as field of view or energy, change. We simply need to reproduce the projector, enabling our method to generalize better than previous feedforward techniques. While handling unknown projection functions can be challenging in this unsupervised approach, these functions can be parameterized, allowing us to estimate the geometry and accurately reproduce projections and calibrations. The geometrical and physical calibration of real biplanar systems is further discussed in the final chapter 5.

Ill-Posedness, Learned Priors and Uncertainty. A key contribution of our work is demonstrating that the 3D reconstruction problem, when constrained by real human anatomies, may not be as ill-posed as traditionally thought. The solution space is restricted by anatomical plausibility, meaning there are relatively few valid 3D anatomies for a given set of X-ray projections. By incorporating learned anatomical priors, the model drastically reduces ambiguity.

We significantly reduced the ill-posedness of the problem, especially in cases where only two projections are available. When the representation error is small—such as when the generator can accurately reconstruct the anatomy—we found in short studies that two projections were sufficient to recover a close 3D structure. This suggests that two X-rays may be sufficient to effectively constrain the anatomical space, indicating a limited number of possible solutions—at least at a primary structural level. Although we lack formal theoretical guarantees for retrieval—as discussed in [Bora, 2017]—experimental results have shown that despite the highly non-convex nature of the problem, the model can still recover solutions close to the ground truth with very few projections. While the representation error is not perfect and still has room for improvement, most of the error stems from out-of-distribution cases.

The success of this approach depends indeed heavily on accurately learning the distribution of anatomical structures. Generative models like VAEs and GANs have shown strong performance in approximating complex distributions, with newer diffusion models offering even greater potential [Sun, 2024b]. Leveraging larger models that capture finer details and greater anatomical variation could significantly improve reconstruction accuracy. Future work should focus on refining these 3D generative models and expanding the dataset to further constrain the solution space. As these models evolve and data availability increases, the range of possible solutions could be further narrowed, enhancing both accuracy and clinical applicability.

To inform practitioners effectively, it is essential to assess the range of plausible reconstructions that match the same X-ray projections, offering insights into both the model's variability and reliability. It will also clarify the extent to which two projections can distinguish details within the learned anatomical manifold.

Rather than relying on a single outcome, generating multiple solutions from the same projections highlights potential ambiguities. Initial reconstructions from the mean latent space produced consistent results; however, using random starting points in the latent space (as in [Bhadra, 2022]) can reveal a diverse set of plausible reconstructions by varying initializations and optimization paths. Similarly, variational inference (e.g., [Marinescu, 2020]) enables exploration of uncertainty by sampling across a distribution of possible outcomes, providing a more comprehensive view of possible anatomical variations.

For clinical translation, it's crucial to ensure this uncertainty remains within clinically acceptable limits— after the model's accuracy has been thoroughly confirmed. Aligning both model accuracy and uncertainty with clinical standards, and evaluating their combined impact on treatment outcomes, will enable practitioners to use the method

confidently if it proves robust and reliable in practical applications.

Learning Biases with Population-Based Priors Relying on population-based priors can introduce learning biases, as discussed previously for CSGM. This reliance may lead to coarse reconstructions for outliers or rare conditions that are underrepresented in the training data. Some failure cases, shown in Figure 3.29 in the Appendix, typically arise from rare anatomies, significant overweight, abnormal postures, or other abnormalities. Such errors may also reflect biases related to specific population characteristics, such as gender or race. Addressing these biases is essential to develop a robust and generalizable model.

To mitigate these effects, a larger dataset or targeted priors for abnormalities could be beneficial, especially with advanced generative models like diffusion models, as well as debiasing techniques. As learning progresses, the quality of manifold approximation should improve, capturing a broader range of anatomical diversity. Additional validation on larger datasets is also necessary to ensure comprehensive population representation.

Resolution and Size We should aim to increase the resolution and volume size to capture more details and a higher degree of anatomical variation. For this, we can either scale up GPU resources or adapt the learning strategy. This could involve learning only the necessary features in latent space, as used with VQ-VAE [Van Den Oord, 2017] in latent diffusion models [Rombach, 2022], or employing neural implicit representations [Mildenhall, 2021], or cascade low-to high res [Sun, 2022] to upsample to full scale.

Furthermore, refinement models could be used to enhance the reconstruction quality. After obtaining an initial coarse reconstruction, a 3D model could be employed to refine both the resolution and the quality, removing artifacts and introducing finer details.

Number of Projections. Determining the optimal number of projections—whether 2, 4, or 8—is crucial for balancing robustness and practicality. As the number of projections increases, reconstruction quality improves, and uncertainty decreases. With only one projection, ambiguity is very high, but with two, important depth ambiguities are resolved. 4 or 8 projections further enhance accuracy, by reducing ambiguity. However, the greatest clinical advantage lies in using just one or two projections, as these don't require patient rotation.

With additional projections, the dependence on strong priors—essential when projections are limited—reduces, as the projections themselves provide sufficient structural constraints. Iterative methods, like NeRF, become more adaptable with more projections, as they rely less on restrictive priors. However, there is a plateau effect: once the projection frequency matches the resolution needed for accurate tissue reconstruction, additional projections provide diminishing returns.

Optimization. We initially used 100 steps for optimization, offering a good balance between performance and speed. However, experiments indicate incremental improvements up to 350 steps, with Adam proving robust in practice. Other optimizers or learning rate schedulers could potentially enhance convergence, and exploring different optimization strategies may further improve both reconstruction quality and speed.

Additionally, understanding the latent space associated with the CT manifold—including its structure (e.g., annular, Gaussian) and the effects of style and disentanglement—can enhance control over the generation and refinement processes. This insight could allow us to better tailor search strategies. Studies such as [Menon, 2020] and [Bhadra, 2022] show that such exploration can significantly improve model performance and robustness.

Computation Time Our method is already quite fast, achieving results in approximately 25 seconds. For applications like pre-treatment positioning, reducing this time to under one minute is a significant improvement over traditional CBCT acquisition. However, for real-time adjustments during treatment for intra-fraction adaptive radiotherapy, further speed improvements are necessary.

To accelerate the process, a more efficient generative model can be developed using techniques like model distillation, GPU splitting, or more advanced optimizer learning schedulers for gradient descent. It is crucial to evaluate the minimum number of iterations required to meet clinical thresholds. One approach is to use encoder methods to quickly estimate the latent space, followed by refined optimization for greater accuracy. Additionally, warm-up techniques—such as using the previous day's reconstruction or the most recent one with a few refinement steps—could further reduce processing time.

3.7 Appendix

3.7.1 2D Experiment : Generation and Reconstruction

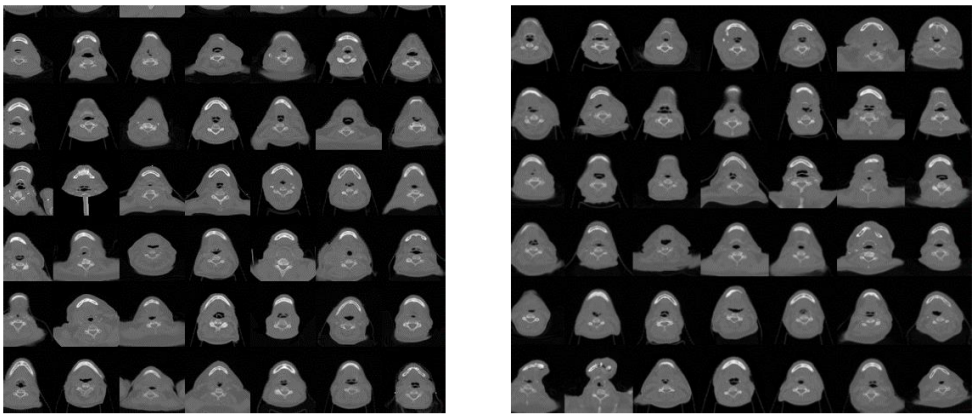

Figure 3.24: Examples of 2D real CTs (left) compared to randomly generated 2D volumes by our StyleGAN2 model trained on 2D slices from our dataset (right), shown on axial slices. It demonstrates a wide diversity of generated anatomies with a good level of detail.

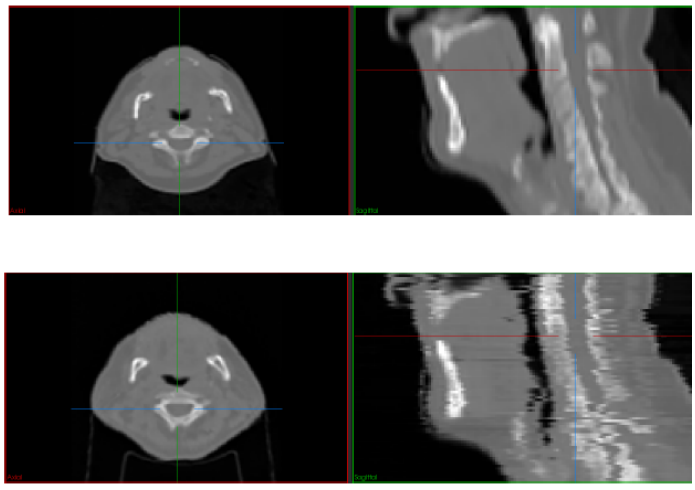

Figure 3.25: Comparison of real ground truth (top) and reconstruction from biplanar X-rays with fan beam geometry (bottom), reconstructed slice by slice. While the overall structure is quite closely retrieved, this highlights inhomogeneities between the independently reconstructed slices.

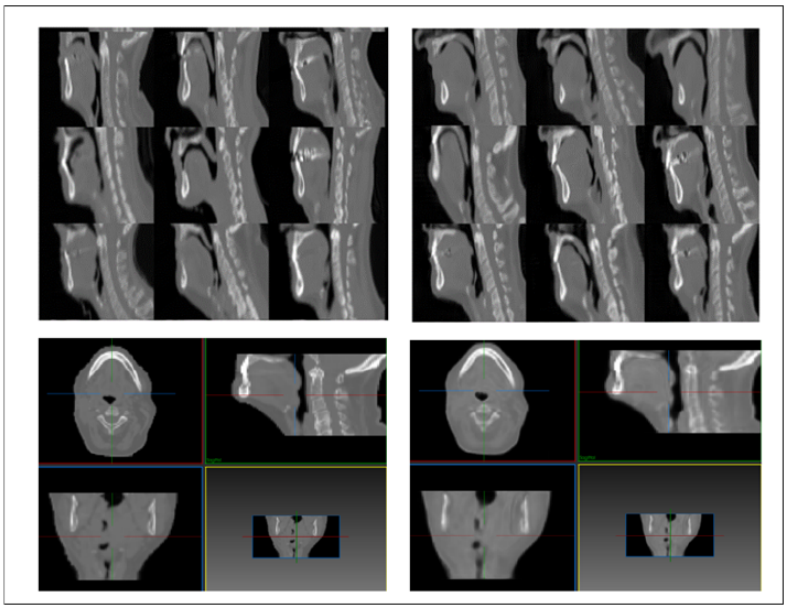

Figure 3.26: Examples of 3D real CTs (top left) vs randomly generated 3D volumes by our model (top right), with view on sagittal slices. Below, recovery by our model (bottom right) of an unseen real 3D CT (bottom left) is represented, with view on axial, coronal and sagittal slices. The recovery well matches the CT structure.

3.7.2 3D Generation

3.7.3 3D Reconstruction

3.7.4 Dosimetry Evaluation

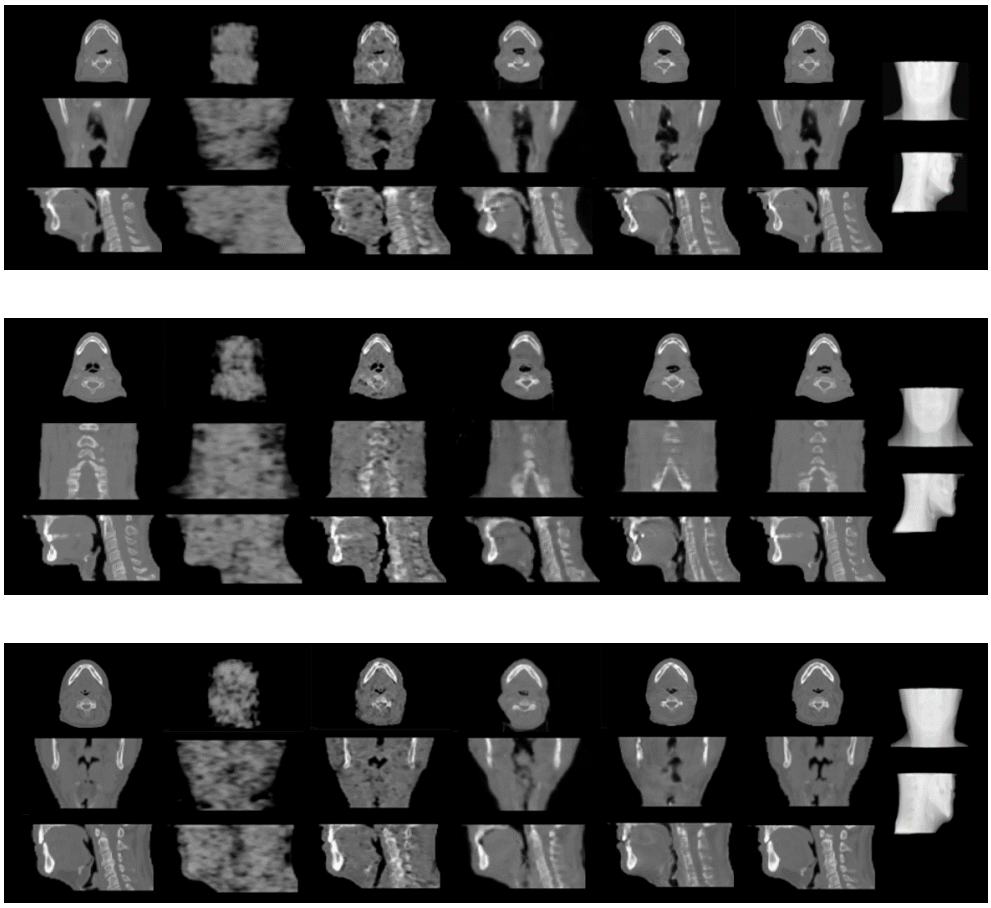

Figure 3.27: **Additional visual comparisons of 3D reconstructions from biplanar projections by our model and baselines.** Our reconstructions are systematically closer to GT and without artefacts.

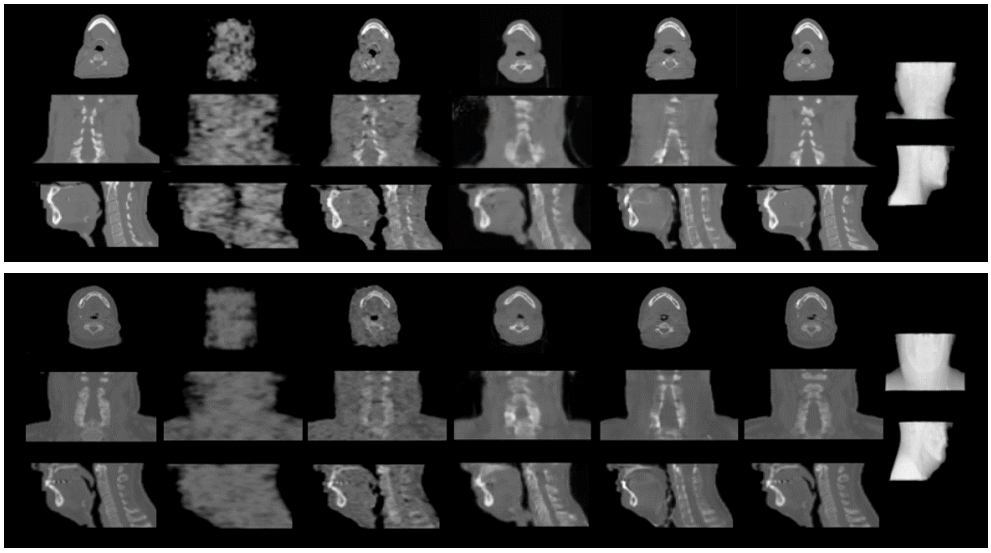

Figure 3.28: **Additional visual comparisons of 3D reconstructions from biplanar projections by our model and baselines.** Our reconstructions are systematically closer to GT and without artefacts.

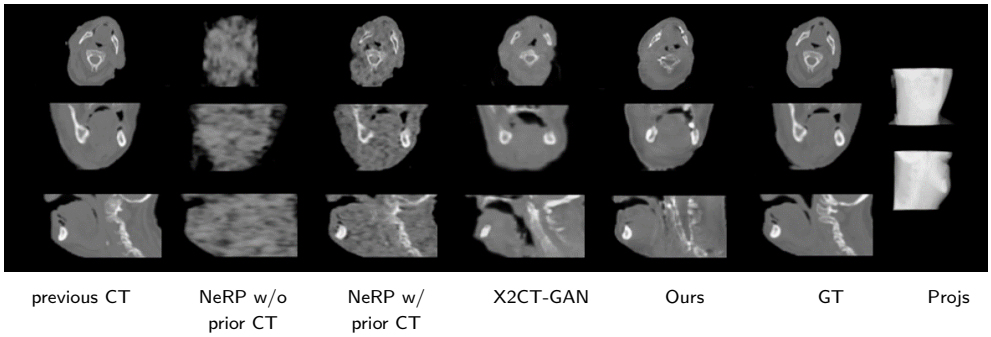

Figure 3.29: **Failure case.** This represents an out-of-distribution case, characterized by substantial overweight and position shift, resulting in visible crane alongside an atypical spine structure, potentially caused by scoliosis.

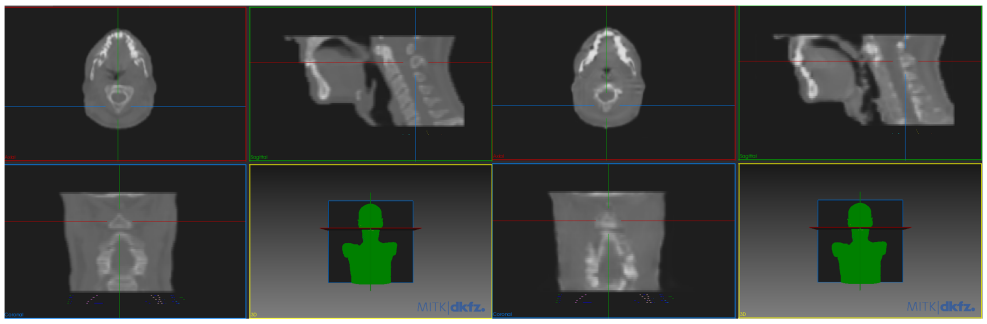

Figure 3.30: Comparison of ground truth (left) and reconstructed image from biplanar X-rays (right), showing good alignment of anatomical structures. This demonstrates the accuracy of the reconstruction method in preserving anatomical details.

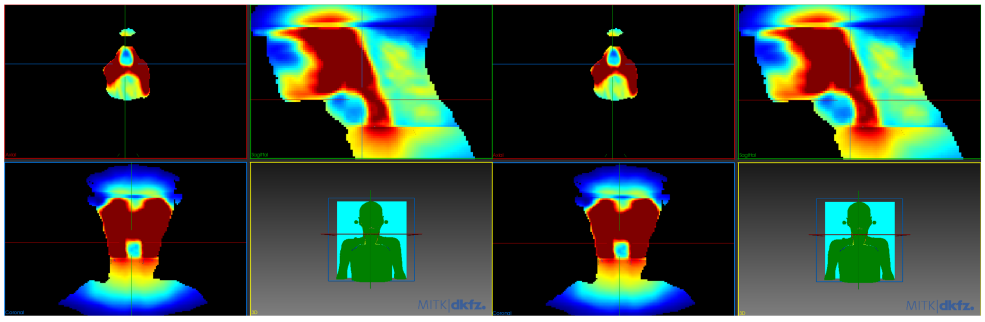

Figure 3.31: Dose simulation on the ground truth (left) and reconstructed image (right right). Similar dose distribution and scatter patterns highlight the anatomical consistency between the reconstruction and actual anatomy.

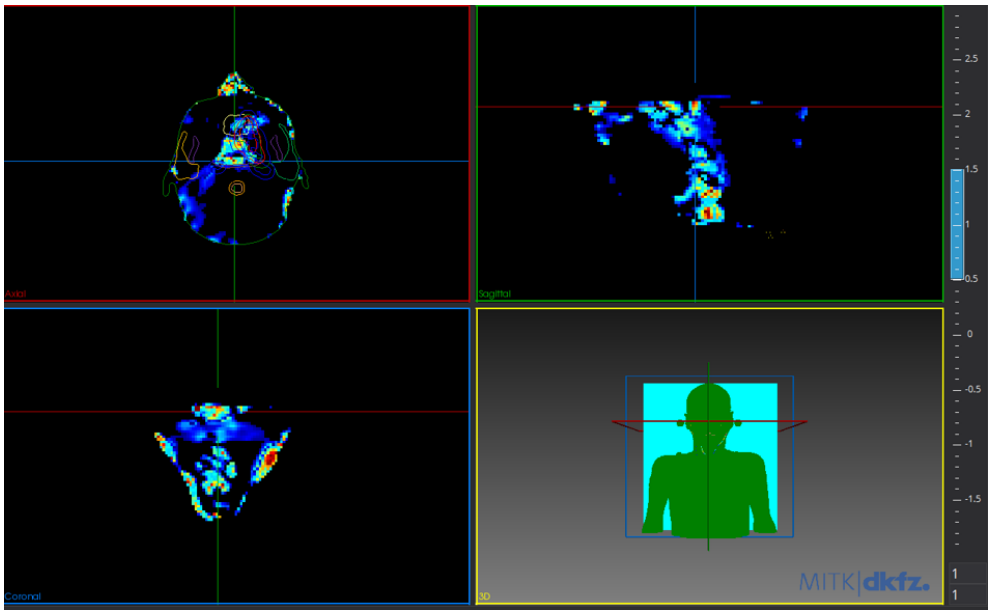

Figure 3.32: Gamma index map with a 3mm distance-to-agreement criterion, illustrating alignment between dose distributions on the ground truth and on the reconstruction. Most points show differences well below 1Gy, with only a few localized areas exhibiting differences greater than 1Gy, indicating a high level of dose distribution agreement.

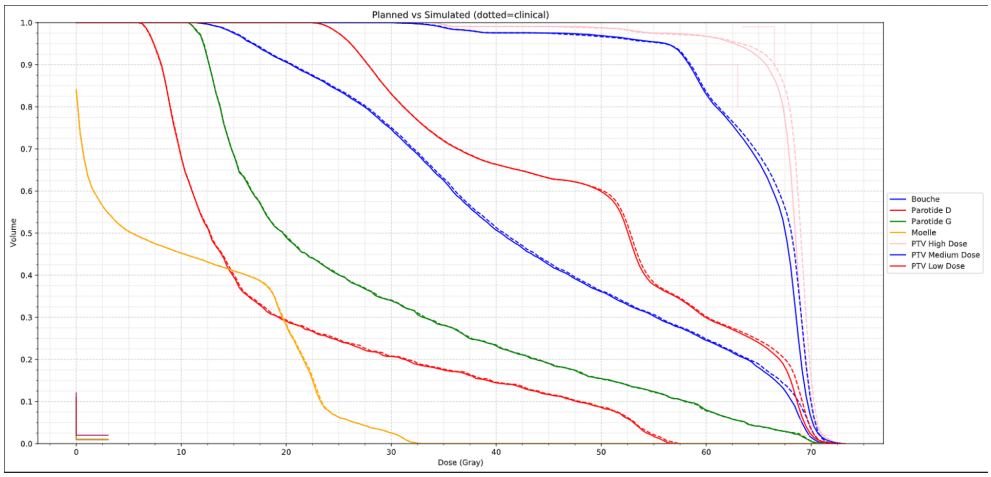

Figure 3.33: Comparison of dose-volume histograms (DVHs) for the simulated doses between the ground truth (planned) and the reconstruction (simulated). The overall alignment is good, with a slight increase observed in high-dose regions for the PTV and mouth. Note: Constraints shown here are not based on actual clinical limits.

Chapter 4

3D Reconstruction-Deformation from Biplanar X-Rays with Pre-Acquired CT

Achieving clinical-grade accuracy and robustness in 3D reconstruction demands methods that are both precise and adaptable. While leveraging generative models hold great potential, they can face limitations in capturing patient-specific nuances and unique anatomical variations, which are essential for clinical translation.

In medical imaging, pre-acquired scans like planning CTs or MRIs are often available, especially in radiotherapy where patients undergo multiple imaging sessions. These prior images provide valuable anatomical references that enhance reconstruction quality by incorporating patient-specific details, crucial when using sparse data from biplanar X-rays. Leveraging these scans enables a personalized 3D model that adapts to treatment-induced changes, such as deformations or weight loss, improving accuracy and reducing the risk of outliers outside the training data distribution.

We propose **XSynthMorph**, an unsupervised method that integrates a pre-captured CT and combines deformation priors with generative models to enhance 3D reconstruction from biplanar X-rays. Unlike previous methods, which struggle with underconstrained deformable alignment, XSynthMorph uses a generative model to guide deformations, adapting precisely to each patient's unique anatomy. To recover a 3D volume that aligns with the two projections, we optimize latent vectors to generate the best 3D volume so that the pre-captured volume well deforms on it to match the projections.

This chapter explores the theoretical foundations of incorporating pre-acquired data to solve inverse ill-posed problems, details the XSynthMorph approach, and presents experimental results that demonstrate its effectiveness compared with previous state of the art and towards adaptive radiotherapy, including rigid registration and deformation of critical

organs.

Contents

4.1 Theoretical Foundations

There are two main approaches to incorporating pre-captured volumes into inverse ill-posed problems to address the sparsity of subsequent measurements: prior-image constrained compressed sensing and registration.

4.1.1 Prior Image-Constrained Compressed Sensing

Compressed sensing, discussed in Section 3.2.2, is a framework used for reconstructing signals from highly undersampled data. While CS has achieved significant success in various fields, its application in medical imaging, such as repeated CT or MRI scans, can benefit greatly from incorporating a pre-acquired image of the same object. This patient-specific information can greatly improve the reconstruction results. However, integrating this data into the process can be challenging.

Prior Image-Constrained Compressed Sensing (PICCS) [Chen, 2008a] leverages prior knowledge from a pre-captured volume to improve the reconstruction of new images from undersampled data, especially for dynamic objects like evolving tumors. Originally introduced for CT imaging, PICCS assumes that while the current and prior images are similar, their differences are sparse under certain transformations (e.g., wavelets or finite differences). This sparsity enables more accurate reconstruction with fewer measurements.

It addresses the following optimization problem, inspired by LASSO formulation 3.15, which incorporates a prior image:

$$\hat{x} = \arg\min_{x} \|y - Ax\|_2^2 + \lambda \left(\alpha \|\Psi(x - x^{(PI)})\|_1 + (1-\alpha)\|\Phi x\|_1 \right), \tag{4.1}$$

where y is the measurement data, A is the forward operator, $x^{(PI)}$ is the prior image, Ψ is a sparsifying transform applied to the difference between the reconstructed image x and the prior image $x^{(PI)}$, Φ is another sparsifying operator applied to the current image x, and α is a weighting parameter that balances between enforcing sparsity in the difference with the prior image and enforcing sparsity in the current image itself.

In this formulation, the optimization aims to enforce sparsity in both the difference between the current image x and the prior image, and the current image itself.

However, this assumption of sparsity in the difference may not always hold. In real-world applications, images are often better modeled as compressible rather than strictly sparse, and their differences may exhibit complex patterns that simple sparsity constraints cannot fully capture [Kelkar, 2021a].

To address these limitations, [Weizman, 2016] introduced adaptive weighting between the prior image and the current image estimate. This approach dynamically adjusts the influence of the prior image on the current reconstruction, providing better results in cases where significant differences exist between the two images.

PICCS with Generative Model Following the same trend as in compressed sensing, deep learning with generative models has been introduced to further improve reconstruction quality beyond traditional sparsity-based methods. These models map both the prior and current images into a high-dimensional latent space, enabling more refined control over the reconstruction process from incomplete measurements. Recent advancements, such as Style-Based Generative Models for Prior Image-Constrained Reconstruction (PICGM) [Kelkar, 2021a], provide an alternative to sparsity-based approaches by leveraging generative models to capture more complex image features and enhance reconstruction accuracy.

Kelkar et al. proposed using the latent representation of a prior image to regularize the reconstruction of a new image. In this approach, the optimization is performed in the latent space of StyleGAN [Karras, 2019]. The prior image constrains certain features, such as shape or texture, while allowing changes in areas like tumors or anatomical shifts. By constraining the latent variables, the reconstruction preserves the overall structure from the prior image (captured by low-level styles) while allowing localized changes (captured by high-level styles).

The objective function is defined as:

$$\hat{w} = \arg\min_{w} \|y - AG(w)\|_2^2 + \lambda \|w - w^{(PI)}\|_{\Sigma}^2, \tag{4.2}$$

where w is the latent vector of the current image, $w^{(PI)}$ is the latent vector of the prior image, and Σ is a covariance matrix that regularizes the difference between the latent vectors.

This approach improves stability and accuracy over traditional compressed sensing methods and effectively handles the challenges of modeling complex differences between images, even with sparse measurements. It leverages the strength of generative models in synthesizing highly realistic images from latent variables.

It has demonstrated superior performance in medical imaging tasks where traditional compressed sensing methods often fall short, particularly in dealing with complex and evolving structures, such as tumors.

Figure 4.1 shows visual results from [Kelkar, 2021a], illustrating MRI reconstruction using a pre-acquired MRI as the prior and sparse measurements from the current day's MRI scan.

While PICCS is effective for scenarios involving minimal changes by incorporating prior images to constrain sparse data, deformable registration can be better suited to handling larger and more complex anatomical shifts. It provides a more robust, well-posed and adaptable framework for integrating prior data, especially when managing substantial deformations in sparse measurements.

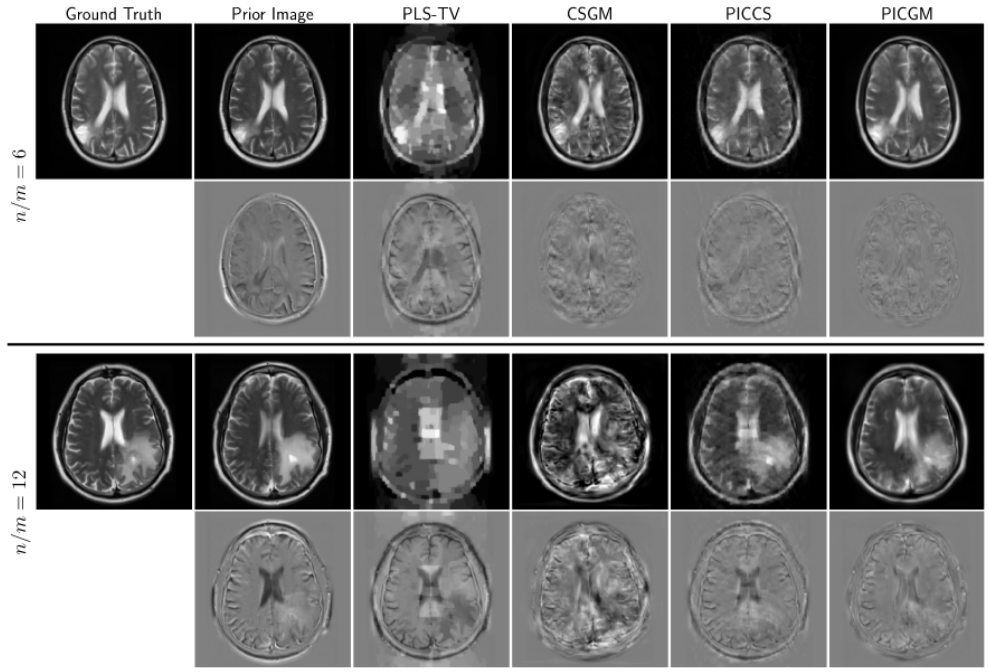

Figure 4.1: Ground truth, prior image, and images reconstructed by PICGM [Kelkar, 2021a] from simulated MRI measurements with varying sparsity ratios of n/m = 6 and n/m = 12, along with difference images for the MRI study.

4.1.2 Registration for Ill-Posed Inverse Problems

Registration is widely used to address inverse ill-posed problems involving sparse measurements by integrating information from multiple time points. In medical imaging, it enables the alignment of pre-acquired high-resolution volumes (e.g., CT or MRI) with sparse or lower-resolution modalities (e.g., X-rays, ultrasound, or 2D slices), enhancing reconstruction quality, compensating for anatomical changes, and integrating data across multiple imaging sessions.

Registration methods can be broadly categorized into rigid and deformable. Rigid registration aligns images through simple transformations like translation and rotation, assuming that the structures do not change shape significantly. This approach works well for cases where minimal anatomical deformation is expected, such as aligning bone structures or other fixed anatomical features.

In contrast, deformable image registration (DIR) is necessary when dealing with more complex scenarios, such as soft tissues or evolving pathologies, where structures can change shape, shift, or deform over time. This method can account for non-linear anatomical changes, making it essential for tracking tumor growth, organ motion, or post-surgical tissue adjustments. Deformable registration offers the flexibility needed to align images in these dynamic contexts, ensuring accuracy in applications such as radiotherapy, where precise alignment of evolving anatomical structures is critical.

Domain Translation for Multimodal Registration Multimodal registration is typically needed to align a high-quality pre-acquired 3D image with subsequent sparse measurements. This process, which involves aligning images from different modalities like X-rays, MRI, and CT, is challenging due to the unique ways each modality represents anatomical structures.

To address these challenges, robust similarity metrics, such as normalized cross-correlation (NCC) and mutual information (MI), are often used, especially when both modalities capture overlapping structural features. Notably, Wells III et al. [Wells III, 1996] introduced mutual information as an important metric for multimodal registration, effectively optimizing the shared information between images to improve alignment across modalities such as MRI, CT, and PET.

However, when direct comparison between modalities is difficult—such as when structures are differently embedded, compressed, or projected in 2D from 3D, or vary in their level of detail—transforming one modality into a common domain can facilitate the registration process. For this deep learning has shown to be effective. For instance, translation techniques using Conditional GAN like Pix2Pix [Isola, 2017] or CycleGANs [Zhu, 2017] have been used to map one domain into another. This approach brings the modalities into a shared domain, improving feature matching and overall registration accuracy.

For example, SymReg-GAN [Zheng, 2021], introduced by Zheng et al., advances the field by employing symmetric GAN-based transformations to ensure consistency in both directions, thereby enhancing the robustness and accuracy of multimodal image alignment. Building on this, Pielawski et al. introduced CoMIR (Contrastive Multimodal Image Representation) [Pielawski, 2020], which leverages contrastive coding to learn shared dense representations, facilitating robust multimodal image registration. Similarly, Casamitjana et al. proposed Synth-by-Reg (SBR) [Casamitjana, 2021], a synthesis-based registration framework that utilizes contrastive learning to improve nonlinear inter-modality registration through synthesis-driven alignment. In a related way, Liu et al. developed a geometry-consistent adversarial model [Liu, 2023] aimed at enhancing unsupervised multimodal deformable image registration by maintaining structural consistency across modalities. Moreover, Tanner et al. applied a GAN to deformable image registration between MR and CT [Tanner, 2018], effectively addressing structural differences and improving alignment accuracy. Another advanced approach in this domain is StructuRegNet [Leroy, 2023a], which integrates CycleGAN [Zhu, 2017] for modality translation between CT and histopathology slices with a deformable registration framework.

In 2D-3D registration, the challenge is not just translation but also bridging the difference in dimensionality. Creating a pseudo-3D reconstruction from 2D X-rays can help facilitate alignment with 3D volumes like CT. By transforming X-rays into a 3D space, the relationship between the X-ray and CT is more accurately captured, reducing the degrees of freedom and improving registration accuracy. This approach is particularly effective for deformable registration with biplanar X-rays, as we will explore in the next section 4.3.2. For deformable registration, there are two types: 3D/3D and 2D/3D. Both methods can

inform sparse measurements using prior scans.

3D/3D Registration

3D/3D deformable registration, which aligns pairs of 3D images, is commonly used to correct artifacts caused by limited measurements by leveraging prior scans. For example, in CBCT-guided head and neck procedures, the lower image quality of CBCT compared to CT, due to artifacts and noise, can hinder accurate treatment planning. Demons deformable registration [Thirion, 1998] is frequently applied to align CT with CBCT, compensating for anatomical changes and reducing the impact of artifacts. The Demons algorithm, introduced by Thirion, effectively handles complex soft tissue deformations [Mencarelli, 2014; Rigaud, 2015]. Studies have shown that deformable registration improves registration accuracy and precision in the presence of tumor changes, which is critical for adaptive radiation therapy [Zhang, 2018; Fortunati, 2014]. However, DIR's performance can be influenced by the choice of algorithm and the presence of artifacts in CBCT [Veiga, 2015].

2D/3D Registration

A more complex case of deformable registration is the alignment of 2D slices or projections with pre-acquired 3D volumes, commonly referred to as slice-to-volume or 2D-3D registration. This technique is particularly useful in scenarios like image-guided interventions or radiation therapy for volumetric reconstruction, such as with biplanar X-rays. Slice-to-volume registration allows physicians to navigate high-resolution pre-operative 3D data using sparse, real-time 2D slices from modalities like ultrasound or X-rays. Unlike traditional 3D-3D registration, this approach must handle significant discrepancies in dimensionality and imaging modality, requiring sophisticated deformation models and regularization to align the datasets effectively.

Several comprehensive surveys provide detailed analyses of 3D/2D registration methods and slice-to-volume registration. For example, Ferrante et al. [Ferrante, 2017] provided an extensive survey on slice-to-volume registration methods. This review presents a complete analysis of the algorithms used to align 2D slices (such as histopathology or ultrasound) with pre-acquired 3D volumes, highlighting the advantages and challenges of various approaches.

MR to Ultrasound Registration One common application of multimodal deformable registration is the alignment of pre-acquired MRI volumes with real-time intraoperative ultrasound images. Since MRI provides highly detailed anatomical information, it is frequently used as a reference to guide the registration of ultrasound images, which are more adaptable for real-time applications. Just to name a few, [Penney, 2004] successfully registered freehand 3D ultrasound with MRI for liver imaging using a vessel-based non-rigid registration approach. In cardiac imaging, [Huang, 2005] achieved dynamic 3D ultrasound to MRI registration for the beating heart, adapting to real-time movements and anatomical changes.

3D to Histology Registration Several methods have been also developed to improve the registration process between 3D and histology. For example, Rusu et al. [Rusu, 2020] proposed a registration approach, which combines multimodal imaging data to enhance tumor localization in the prostate. This framework integrates histopathological data with pre-surgical MRI, facilitating accurate delineation of tumor boundaries.

Similarly, Ohnishi et al. [Ohnishi, 2016] investigated deformable image registration to align pathological images with MR images, showing significant improvements in spatial correspondence and clinical applicability. They highlighted the need to account for anatomical deformations that occur during histological preparation.

Li et al. [Li, 2017] introduced a method for co-registering ex vivo surgical histopathology with in vivo MRI of the prostate. Their approach employed multi-scale spectral embedding to enhance registration accuracy, allowing for a better understanding of the spatial relationship between histological features and imaging signals.

Additionally, StructuRegNet [Leroy, 2023a] aligns 3D CT scans with 2D histopathology slides. It utilizes adversarial modality translation using CycleGAN to merge the two modalities into a shared domain, improving alignment without the need for full 3D reconstruction. This method is particularly effective for tumor mapping in head and neck cancer, enabling precise pixel-wise registration between CT and histopathology data towards virtual biopsy.

CT to X-ray Registration Aligning pre-operative 3D CT volumes with intraoperative 2D X-ray images is a critical task in image-guided interventions, but it is inherently ill-posed due to the ambiguity of X-ray images. The registration process typically starts with rigid alignment, where a transformation is applied to match CT data with X-rays based on anatomical landmarks.

Markelj et al. conducted a comprehensive review of 3D/2D registration methods for image-guided interventions, focusing on aligning pre-acquired 3D volumetric data with intraoperative 2D fluoroscopic X-ray images. The review covers key aspects such as image modality, dimensionality, geometric transformations, and optimization procedures, providing a detailed overview of techniques for 3D/2D alignment in clinical contexts.

A major finding was the importance of dimensional correspondence, where strategies

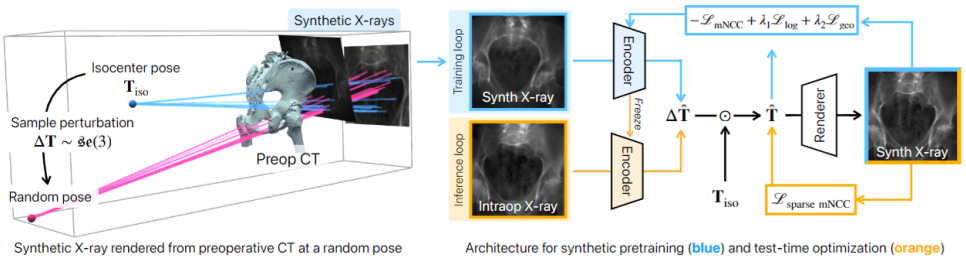

Figure 4.2: DiffPose method [Gopalakrishnan, 2024]. Left: Random perturbations generate camera poses around the isocenter T_{iso}. Right: An encoder predicts the pose of a synthetic X-ray with image similarity and $SE(3)$-geodesic losses. During inference, the encoder estimates the pose of a real X-ray, refined via test-time optimization and differentiable rendering.

like projection, back-projection, and reconstruction were evaluated based on clinical needs. The review highlighted that intensity-based methods, particularly those using DRRs from CT images, are the most common for 3D/2D registration. Though efficient, DRR-based methods face challenges like high computational costs and limited capture range, making precise initial alignment crucial.

Feature-based methods were also examined, relying on anatomical points, curves, or surfaces for registration. These methods offer faster processing but depend heavily on accurate feature segmentation, which can be challenging.

The review further explored optimization techniques, noting the importance of non-rigid transformations for capturing time-related anatomical changes, especially in soft tissues. It compared extrinsic methods, which use markers for fast and accurate but invasive registration, with non-invasive intrinsic methods, which rely on effective feature extraction for reliability.

The main methods in CT to X-ray registration include Tomazevic et al. [Tomazevic, 2003], who developed a method utilizing bone surface normals from CT and gradients from X-rays, achieving high accuracy with for lumbar vertebrae. Similarly, Livyatan et al. [Livyatan, 2003] presented a gradient-based rigid registration method that optimizes both speed and accuracy by using volume gradients to manage outliers, resulting in low target registration errors.

Zollei et al. [Zollei, 2001b] introduced a mutual information-based registration approach, enhancing alignment between CT and fluoroscopic X-rays through a sparse histogramming method, which leads to robust and efficient registration. Recent advancements in deep learning have further improved CT to X-ray registration. Jaganathan et al. [Jaganathan, 2023] proposed a self-supervised framework that eliminates the need for paired annotated datasets by combining simulated training with domain adaptation, achieving a registration accuracy of around 1.83 mm on real X-ray images.

Additionally, Esteban et al. [Esteban, 2019] developed a fully automatic X-ray to CT registration system that employs deep learning for pose initialization and refinement, achieving high accuracy. More recently, Zhang et al. [Zhang, 2023a] proposed a patient-

specific, self-supervised registration framework that accurately estimates X-ray pose using simulated patient-specific X-rays, reaching a mean projection distance of 1.55 mm on real X-ray images. Also, Gopalakrishnan et al. [Gopalakrishnan, 2024] introduced DiffPose, a self-supervised approach that leverages differentiable physics-based projector to perform accurate 2D/3D rigid registration. By training a CNN on synthetic X-rays rendered from a preoperative CT, DiffPose achieves sub-millimeter accuracy in surgical datasets, significantly outperforming existing unsupervised methods and even surpassing some supervised baselines. Figure 4.2 illustrates the DiffPose method.

Research and clinical applications have traditionally focused on rigid registration. Deformable registration, however, is inherently more complex, especially when working with only one or two projections. Advances in deep learning have made deformable registration more feasible and faster by using learned priors. Methods like the CNN-based approach introduced by Lecomte et al. [Lecomte, 2022] for real-time 2D-3D deformable registration using a single X-ray projection, along with other methods we'll explore in the next section, show promise but still face limitations. As we will demonstrate, these methods remain under-constrained and offer room for further improvement.

4.2 3D Reconstruction-Deformation from Biplanar X-Rays with Pre-Acquired CT

Returning to the challenge of 3D reconstruction from biplanar X-rays, we introduce a novel unsupervised approach that recovers and registers a 3D volume using only two planar projections, leveraging a pre-acquired 3D volume, specifically the planning CT used for radiotherapy.

Figure 4.4 illustrates existing approaches to decreasing the number of projections when reconstructing or registering a volume using only very few X-rays.

As stated in previous chapter 3, one approach is to train a deep model to regress the volume from a number of projections in a supervised way [Henzler, 2018; Shen, 2019; Ying, 2019; Jiang, 2022; Tan, 2022; Lu, 2022; Tan, 2023; Zhang, 2023b; Shen, 2022b; Zhang, 2021a; Tian, 2022; Wang, 2023], but such direct inference often results in limited reconstruction.

Instead, our previous approach X2Vision learns an anatomical prior and optimizes its parameters to match the projections. Such optimization approaches (see also [Shen, 2022a]) tend to generalize better and bring more realistic results.

As said in the introduction, in modern medical practice, CT and MRI scans are now widely used for treatment planning and diagnostics. This provides a volume captured under a different patient pose and different from the volume to reconstruct by medically-relevant changes such as weight loss or tumor transformation. X2Vision ignores this pre-captured volume but 2D3DNR [Dong, 2023] deforms it given several projections.

How can we exploit both sources of information, anatomy knowledge from a generative

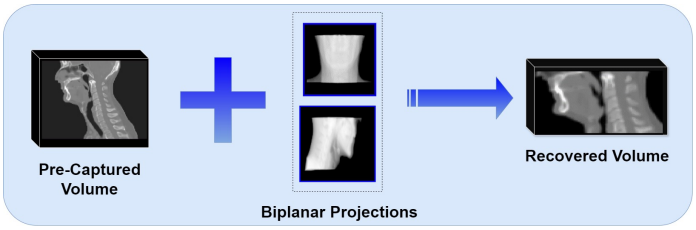

Figure 4.3: The goal of our method is to recover an accurate 3D volume given two projections of the patient and a volume acquired at the beginning of the therapy. As discussed in the introduction, this ability unlocks better therapy procedures. Combining correctly these sources of information is however challenging.

model and the pre-acquired volume, to improve the reconstruction quality to the point it can actually be used for medical applications? This is the goal of our method illustrated in Figure 4.3.

Combining the strengths of approaches like X2Vision and 2D3DNR is not straightforward, as they work in fundamentally different ways. We begin by noting that when 2D3DNR deforms the pre-acquired volume using very few projections, unrealistic deformations may occur. Additionally, some complex anatomical variations may not be well-represented in the dataset, limiting the model's ability to generalize effectively.

To address this, we propose guiding the deformations using a volume generative model. Note that this is different from X2Vision, which directly optimizes the generative model parameters to match the projections: Our approach allows to deform the pre-acquired volume under the guarantee that the resulting volume is anatomically possible.

To do so, we optimize over the generative model parameters so that the pre-acquired model matches well the projections after being deformed onto the generated volume. This is illustrated in Figure 4.4. Compared to 2D3DNR for example, we guarantee that the deformed pre-acquired volume is anatomically possible, since it is constrained to be close to a generated (thus anatomically correct) volume. Moreover, we also have the guarantee that the deformed volume well matches the projections. Compared to X2Vision, because our approach predicts the pre-acquired volume after deformation, it captures the patient's unique anatomy or abnormalities accurately. This is by contrast with a generated volume, as done by X2Vision, which can lack details.

While predictable deformations, such as lung movement in 2D3DNR, can be captured effectively, more complex factors—like radiotherapy-induced changes (e.g., weight loss) or the intricate anatomy of the head and neck—are harder to predict. The head and neck region is particularly challenging due to its highly heterogeneous and complex structures, such as the larynx, jaw, and teeth. These deformations involve a combination of neck twisting, jaw articulation, and changes caused by weight loss or tumor growth, making accurate recovery of fine anatomical details especially difficult.

In our evaluations, we continue to focus on head-and-neck CT scans from cancer patients undergoing radiotherapy, using same cohorts from two different medical centers, as in previous evaluations.

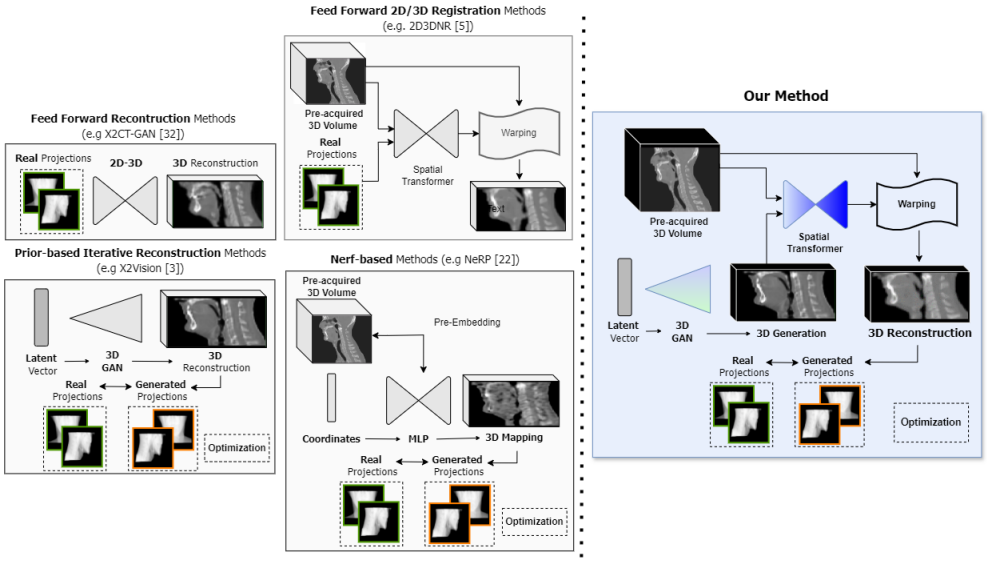

Figure 4.4: **Our method can recover an accurate 3D volume given two projections of the patient and a volume acquired at the beginning of the therapy**, to unlock better therapy procedures. Combining correctly these sources of information is however challenging: **(1)** Feed-forward reconstruction [Henzler, 2018; Shen, 2019; Ying, 2019; Jiang, 2022; Tan, 2022; Lu, 2022; Tan, 2023; Zhang, 2023b; Shen, 2022b; Wang, 2023], as introduced in previous chapter 3 directly predict a volume from a set of projections. **(2)** Methods such as 2D3DNR [Dong, 2023] deform a pre-acquired volume of the patient based on input projections. However, the predicted deformation can become under-constrained when the number of projections is very low as two. **(3)** Methods such as our previous method X2Vision [Cafaro, 2023d] first learn a volume generative model and optimize the parameters of this model to match the projections. However, they are not able to exploit the pre-acquired volume. **(4)** NeRF-based methods [Shen, 2022a] can take the pre-acquired volume as input and optimize on the volume to match the projections—however, they do not exploit any anatomical knowledge besides the pre-acquired volume. **(5)** To avoid predicting incorrect deformations when the number of projections gets too low as one or two, we propose to guide the deformations using a volume generative model. Note that this is different from X2Vision, which relies on a generative model to directly create the predicting volume. Instead, our approach deforms the pre-acquired volume under the guarantee that the resulting volume is anatomically plausible.

We compare our method against top-tier techniques, including our previous approach X2Vision [Cafaro, 2023d], 2D3DNR [Dong, 2023], and the NeRF-based approach [Shen, 2022a]. The results demonstrate our method's superiority and ability to capture important anatomical details.

Our method demonstrates high-quality deformable and rigid registration, indicating a move towards more precise biplanar systems over traditional 3D visualization. Unlike typical 2D/3D registrations focused on bones, our method enables finer adjustments due to our detailed 3D reconstructions. Our approach closely matches patient anatomy, potentially enhancing daily treatment precision and enabling daily adaptive radiotherapy, while significantly reducing irradiation.

4.3 Related Work

4.3.1 3D Reconstruction from a Few X-Rays

Previous chapters have highlighted the state of the art in 3D reconstruction from bi-planar X-rays, focusing primarily on feedforward methods [Henzler, 2018; Shen, 2019; Ying, 2019; Jiang, 2022; Tan, 2022; Lu, 2022; Tan, 2023; Zhang, 2023b; Shen, 2022b; Zhang, 2021a; Tian, 2022; Wang, 2023]. Iterative methods, such as those proposed in [Shen, 2022a; Cafaro, 2023d], which optimize the reconstructed volume during inference based on projections, generally achieve better performance. Our previous approach X2Vision [Cafaro, 2023d] leverages a learned 3D manifold to produce realistic and accurate reconstructions, while others rely on NeRF-based approaches [Shen, 2022a; Zha, 2022]. These iterative methods ensure projection consistency during inference and demonstrate superior generalization.

However, almost no solution has introduced the possibility of integrating pre-acquired volume to enhance the reconstruction.

[Shen, 2022a] already exploits the pre-acquired volume but simply by using the pre-acquired volume to initialize the NeRF. When using very few projections, this fails as shown in previous chapter 3.4. Our approach efficiently combines such a pre-acquired volume thanks to a prior on its possible deformation.

4.3.2 2D/3D Deformable Registration from a Few X-Rays

Our approach is now directly related to the 2D/3D deformable registration problem, as we deform the pre-acquired 3D volume by comparing its projections to the captured X-ray projections.

A common approach to 2D/3D deformable image registration involves solving an optimization problem to determine the transformation parameters that best describe the deformation between a 3D volume and a set of 2D projections, as explored in [Flach, 2014; Prümmer, 2006; Tian, 2020; Zikic, 2008]. This process typically relies on measuring image similarity by comparing actual CT projections with corresponding simulated projections (DRRs). However, when only a limited number of projections is available, traditional non-deep learning methods face challenges, as the problem becomes underconstrained.

Some learning-based methods for 2D/3D deformable image registration use feedforward predictions for faster registration [Foote, 2019; Li, 2020; Pei, 2017; Zhang, 2021a]. For instance, [Zhang, 2021a] employs a U-Net architecture to predict deformation fields directly from a CT volume and captured X-ray projections, while maintaining projection consistency during training. This method uses the FDK method [Feldkamp, 1984a] to transform X-rays into a coarse 3D representation, bringing the data into the 3D domain. It also integrates a forward-projection layer to generate DRRs from the deformed CT, which are then compared with the captured X-ray projections to guide the learning process and ensure consistency. Figure 4.5 illustrates this pipeline.

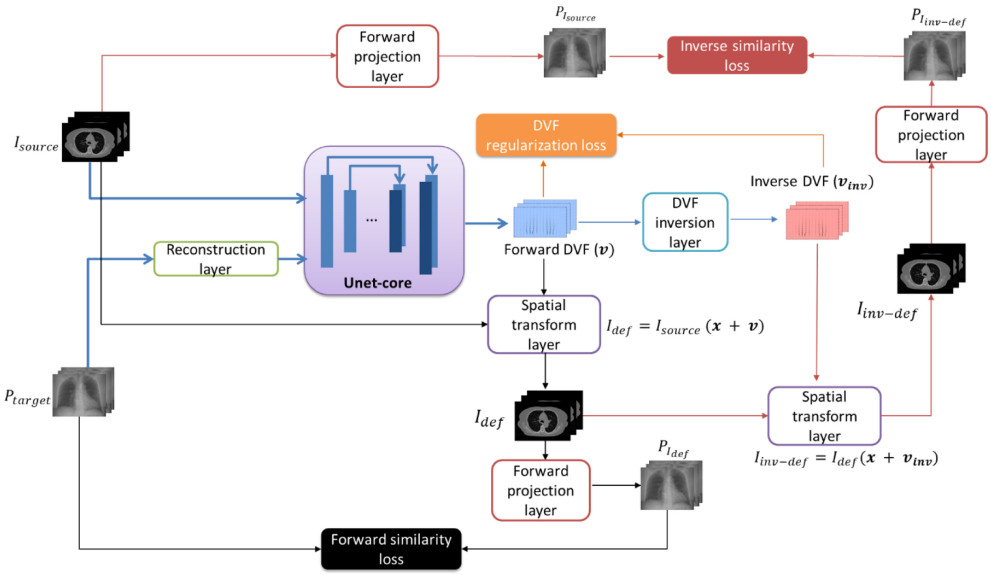

Figure 4.5: The pipeline from [Zhang, 2021a] uses a U-Net architecture to predict deformation fields directly from CT volumes and 2D X-ray projections. The method integrates a forward-projection layer to generate DRRs from the deformed CT, which are compared with the captured X-ray projections to ensure consistency during training.

While learning-based methods for 2D/3D deformable registration show promise, they often struggle with spatial ambiguity inherent in 2D projections. This limitation arises from the lack of 3D spatial information in the loss function, which is crucial for learning accurate 3D deformations, as noted in [Tian, 2022]. Some works, such as those by [Pei, 2017] and [Li, 2020], attempt to address this by using prior data to constrain deformations within a realistic transformation space, typically via principal component analysis (PCA) derived from a cohort of training data. While this reduces the problem's complexity, it does not fully resolve the spatial ambiguity introduced by the 2D measurements.

Also, methods like [Foote, 2019] build subject-specific deformation subspaces by training on 3D image pairs of the same patient, avoiding spatial ambiguity during inference. However, these methods assume access to multiple 3D images of the same patient during training, which is not feasible in most clinical settings where only one CT and a set of X-rays are available. As a result, this approach is limited when generalizing across different patients or imaging scenarios.

The key challenge for 2D/3D DIR is overcoming the spatial ambiguity caused by the limited dimensionality of 2D projections. A promising solution is to use deep learning to incorporate accurate 3D spatial information during training, using high-quality 3D-3D image pairs. This would reduce ambiguity in training while ensuring generalizability at test time, even without 3D image pairs.

The LiftReg method proposed by [Tian, 2022] addresses spatial ambiguity in several ways. First, it extracts 3D spatial information from multi-channel backprojected volumes, rather than relying solely on 2D features. Additionally, it incorporates prior knowledge

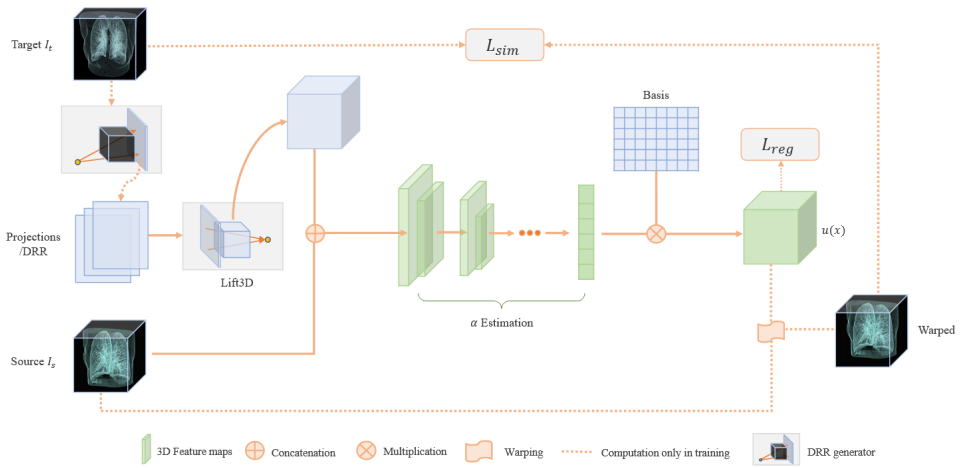

Figure 4.6: LiftReg pipeline [Tian, 2022]. The network extracts 3D spatial information from multi-channel backprojected volumes and uses a PCA-reduced transformation space to manage complex deformations and reduce spatial ambiguity. It includes three modules: Lift3D (for back-projecting 2D images into 3D), α estimation (for predicting deformation vector field coefficients), and a warping module (for deforming the source image I_s with the predicted inverse transformation $\phi^{-1}(x)$). Training flow is shown with an orange dotted line, and learnable components are highlighted in green.

of patient motion through a PCA-reduced transformation space, explaining most of the deformation variability across thousands of cases. This PCA-based subspace simplifies the deformation model while maintaining accuracy. LiftReg's innovation lies in its ability to leverage high-quality 3D-3D pairings during training to guide deformations using only X-rays at inference time, effectively "lifting" 2D data via a 3D context. This approach enhances the model's ability to capture accurate 3D deformations, addressing both the spatial ambiguity and the large degrees of freedom typically encountered in traditional methods that only rely on 2D X-rays. Figure 4.6 illustrates this pipeline.

However, while the use of a PCA-reduced transformation space is helpful, it may not capture the full range of possible deformations, especially for complex anatomical variations, making it somewhat limited. A more advanced approach would be to directly learn the transformation space, enabling more accurate and flexible deformation modeling.

Building on this idea, [Dong, 2023] introduced a method, referred to as 2D3DNR that transitions from 2D biplanar projections into 3D space. This is accomplished by first estimating a 3D feature map from the projections, rather than using simple backprojection. Then a 3D-to-3D deformation learning process using a Attention U-Net-based architecture [Schlemper, 2019]. This method shows promise by leveraging the richer information available in the 3D feature space to enhance registration accuracy.

Our method advances these approaches by incorporating two types of priors: a prior on the predicted 3D volume and a prior on the deformation of the pre-acquired volume. Importantly, these priors are learned in an unsupervised manner, which allows for greater flexibility and adaptability to various clinical scenarios. As demonstrated by our results,

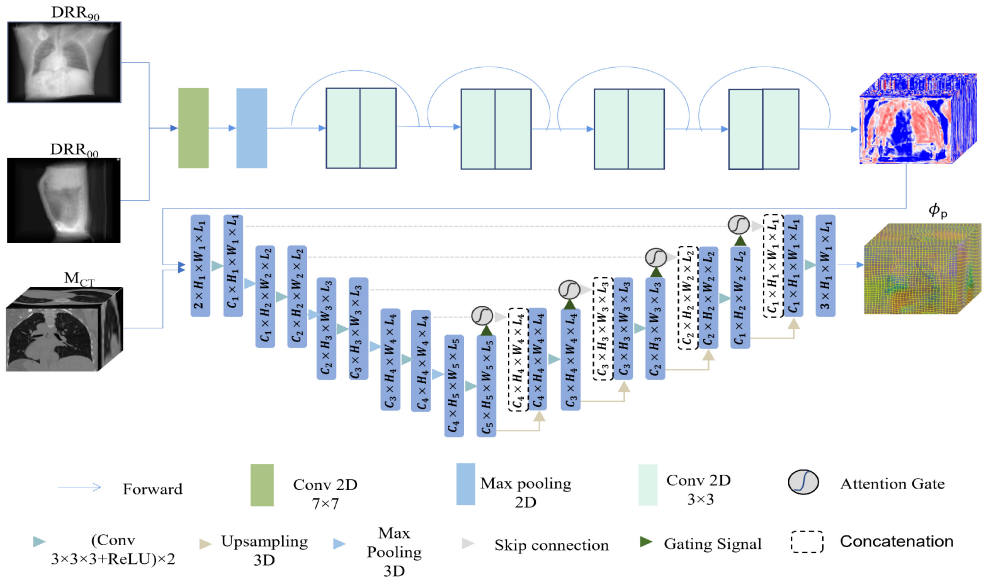

Figure 4.7: 2D3DNR pipeline [Dong, 2023]. First, 2D DRRs at orthogonal angles are processed by residual blocks to obtain 3D feature maps. Then, the feature maps and moving images are fed into a 3D U-Net. The final output of this network is the predicted 3D deformation field.

this dual-prior approach outperforms previous state-of-the-art methods by offering more precise and robust 3D reconstructions and deformations with very few projections.

4.4 XSynthMorph

4.4.1 Problem Formulation

Given a limited set of projections $\{I_i\}_i$, namely two orthogonal planar ones, our objective is to reconstruct the 3D tomographic volume v responsible for these projections.

A previously-captured volume v^- of the patient is available as well. Between v^- and v are both rigid and non-rigid transformations, as well as more complex transformations such as tumor growing or shrinking. We thus seek the transformation of v^- to v.

Finding the deformation is ill-posed when the number of projections becomes small. Our key contribution lies in the following formulation that enforces the predicted deformation of v^- to produce an anatomically correct volume:

$$\mathbf{g}^* = \underset{\mathbf{g}}{\operatorname{argmin}} \sum_i \mathcal{L}_i(S(v^-, v(\mathbf{g})), I_i) + \mathcal{R}(\mathbf{g}) \,. \tag{4.3}$$

We briefly describe below each component of this formulation, then describe them in more details in the rest of the section:

- $v(.)$ is a generative model of volumes of parameters $\mathbf{g}$, i.e., $v(\mathbf{g})$ is a generated volume. In practice, we use a model similar to the one in X2Vision. However, X2Vision uses

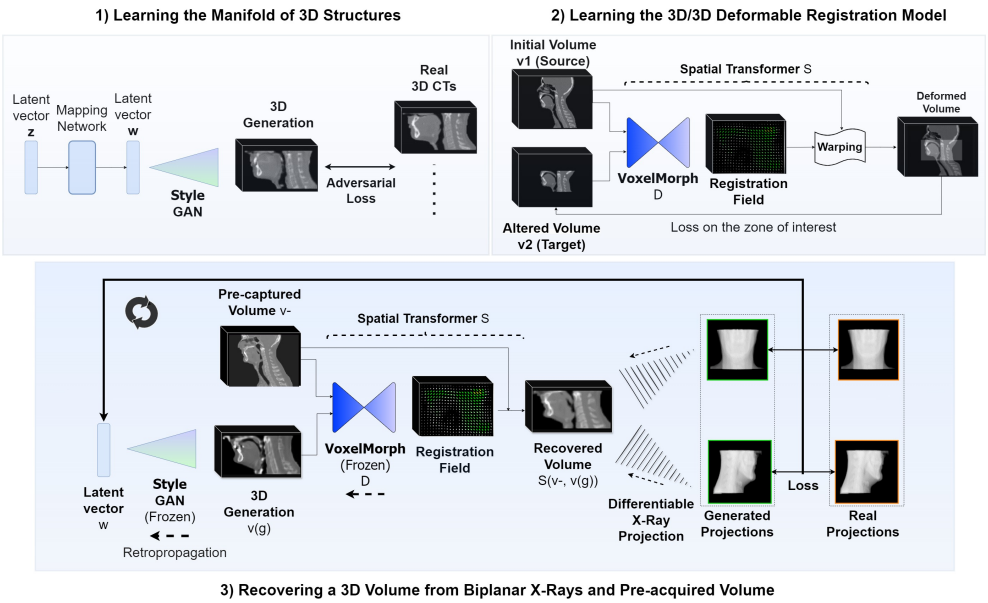

Figure 4.8: **Our pipeline.** We first train a generative model to generate 3D volumes in a low-dimensional manifold and a 3D/3D deformable registration model between two volumes. Now, given two projections and a pre-captured volume of the patient, we recover a 3D volume corresponding to the two projections by finding the latent vectors that generate the best 3D volume so that the pre-captured volume well deforms on it to match the projections. We iteratively refine the generation and deformations based on the discrepancy between the generated and actual projections.

this model to directly predict the final volume v. The key difference in our method is that we use it here to guide the deformation of v^-.

- $S(v^-, v(\mathbf{g}))$ is a spatial transformer [Jaderberg, 2015] [1] trained to predict directly the transformation between two volumes. Here, $S(v^-, v(\mathbf{g}))$ returns volume v^- after deformation to align on the generated volume with the deformation consistent with training data. We use a spatial transformer very close to the one proposed by VoxelMorph [Balakrishnan, 2019].
- $\mathcal{L}_i$ is a loss term that compares the projections of deformed volume $S(v^-, v(\mathbf{g}))$ with the input projections I_i.

Intuitively, the optimization on $\mathbf{g}$ generates a volume $v(\mathbf{g})$ that guides the deformation of v^- thanks to the first term and exploits prior knowledge on volumes to recover. After optimization, our method returns $S(v^-, v(\mathbf{g}))$, the pre-acquired volume after deformation.

Figure 4.8 illustrates our pipeline that implements our approach: We first train generative model $v(\mathbf{g})$ as well as spatial transformer S, both in an unsupervised way. Given two input projections, we then optimize the parameters $\mathbf{g}$ of the generative model, which gives us deformed volume $v^* = S(v^-, v(\mathbf{g}^*))$. We describe the generative model $v(\mathbf{g})$, spatial transformer S, and loss term $\mathcal{L}_i$ below.

[1] Not to be confused with Transformers [Vaswani, 2017].

4.4.2 Generative Model $v(.)$

Similar to X2Vision, we learn the generative model $v(.)$ using GANs, specifically 3D StyleGAN [Hong, 2021]. The model is the same as described in the previous method 3.4.2. We decompose the parameters $\mathbf{g}$ into a latent vector $\mathbf{w}$ and Gaussian noise vectors $\mathbf{n} = \{\mathbf{n}_j\}_j$, such that $\mathbf{g} = [\mathbf{w}, \mathbf{n}]$.

Similarly to X2Vision, to ensure the predicted volume remains within the space of anatomically possible volumes, the regularization term $\mathcal{R}(\mathbf{g})$ is defined as a sum of regularizations on $\mathbf{w}$ and $\mathbf{n}$:

$$\mathcal{R}(\mathbf{g}) = \mathcal{R}(\mathbf{w}, \mathbf{n}) = \lambda_w \mathcal{L}_w(\mathbf{w}) + \lambda_c \mathcal{L}_c(\mathbf{w}) + \lambda_n \mathcal{L}_n(\mathbf{n}) \,. \tag{4.4}$$

The λ_* are fixed weights.

4.4.3 Spatial Transformer S

We use a spatial transformer S similar to the one introduced by VoxelMorph [Balakrishnan, 2019]. It can be decomposed into:

$$S(v_1, v_2) = \mathcal{W}(v_1, D(v_1, v_2)) \,, \tag{4.5}$$

where $D(v_1, v_2)$ is a deep network predicting a deformation field from v_1 to v_2; and $\mathcal{W}(v_1, D(v_1, v_2))$ deforms volume v_1 according to the deformation field predicted by D. Model D is trained to predict deformation $\mathcal{W}$ between two volumes v_1 and v_2 by minimizing $\lambda_s \|v_2 - S(v_1, D(v_1, v_2))\|^2 + \lambda_D \|\nabla D(v_1, v_2))[x]\|^2$, over a training set of corresponding volumes $\{(v_1, v_2)\}$. The second term is a smoothing loss that mitigates sharp local fluctuations and promote smoothness of the predicted field. λ_s and λ_D are balancing weights that adjust the emphasis between similarity and regularization during training.

Maintaining a 1-to-1 mapping in medical image registration is essential to prevent tearing, folding, or overlap during deformation. Inspired by VoxelMorph, our model predicts a velocity field, which is then integrated using the scaling and squaring method [Arsigny, 2006], a common technique in diffeomorphic registration. This ensures the deformation remains smooth, invertible, and free of singularities, preserving anatomical structure and natural tissue movement while avoiding geometric inconsistencies.

4.4.4 Loss Term $\mathcal{L}_i$

As in X2Vision, we take term $\mathcal{L}_i(v, I_i)$ as the weighted sum of the Euclidean distance and the perceptual loss [Johnson, 2016] as we observed that this combination results in the best results. To generate projections, we used a realistic differentiable cone-beam projector as in X2Vision.

4.4.5 Warm-Up

Before optimizing Eq. (3.35) we first retrieve an initial volume estimate $v(\mathbf{g})$ by performing several gradient descent steps of objective

$$\sum_i \mathcal{L}_i(v(\mathbf{g}), I_i) + \mathcal{R}(\mathbf{g}) \,, \tag{4.6}$$

starting from random initialization for $\mathbf{g}$. We use 10 iterations in practice. This provides a better initialization for $\mathbf{g}$ before optimizing Eq. (3.35) and speeds up convergence.

4.5 Experiments

We evaluate our method for our main target application, namely head-and-neck cancer radiotherapy. As mentioned previously, head and neck exhibits many fine details and complex deformations and is representative of many of the different challenges of volume recovery.

In this section, we introduce our dataset, models for learning key priors, and present both quantitative and qualitative comparisons with state-of-the-art methods. We also include an ablation study to evaluate the contribution of our priors. Our method recovers high-quality volumes in only 1 minute. While some other methods are faster, the trade-off fidelity/runtime is well acceptable as clinical CBCT acquisition and FDK reconstruction currently require more than 2 minutes.

4.5.1 Datasets

Volume Generator Learning. We trained our GAN model for $v(\mathbf{g})$ on the same dataset as X2Vision.

Longitudinal Radiotherapy Data. We also used same longitudinal radiotherapy data as in previous method 3.4. With patient consent, we had compiled planning CT scans and subsequent CBCT scans from 242 patients across two medical centers (CLB and IGR), one contributing 177 and the other 65 cases. These datasets, distinct in protocols and scanning equipment, offer a diverse basis for traning and evaluation.

3D/3D Deformable Registration Training. More precisely, to train our 3D/3D deformable registration model, we randomly selected 146 patients for training, 16 for validation, and 10 for testing. We paired each initial CT with any subsequent CT from the same patient to obtain a large training set of more than 1250 pairs.

Volume Recovery. The second part of volumes, used for evaluation, includes 70 patients with the most marked longitudinal alterations. We selected them by comparing their CBCTs with planning CTs. We paired the planning CT with each patient's final vCT(virtual CT as CT deformed on CBCT)—which underscores the utmost discrepancies. We used the planning CT as pre-captured volume. Biplanar projections were synthesized from the last 3D volumes, focusing on the reconstruction area.

4.5.2 Implementation Details

3D/3D Deformable Registration Training. To develop our 3D/3D deformable registration model, we employed the VoxelMorph architecture [Balakrishnan, 2019]. We maintained the channel depths in the encoder at 16-32-32-32, and in the decoder at 32-32-32-32-32-16-16, mirroring the configurations of the original architecture. Initial large CT scans are resized to $96 \times 128 \times 160$ ($2.67 \times 3 \times 3.5mm^3$) for GPU compatibility. Subesequent scans, focusing on the reconstruction area, of shape $80 \times 96 \times 112$ ($1.3 \times 2.4 \times 1.9mm^3$), are downsampled and padded to match initial scan dimensions. Masks are created for targeted training in this region. At test-time, the area of interest is extracted and resized to its original dimensions.

Our loss function weights the mean squared error on the reconstruction area and the gradient regularization term with $\lambda_{sim} = 1$ and $\lambda_{grad} = 1\mathrm{e}{-2}$ respectively. We incorporated 7 integration steps on velocity fields, downsampled by half, to ensure diffeomorphic displacement fields with computational efficiency. Our model, implemented in PyTorch, was optimized using Adam with a learning rate of $1\mathrm{e}{-4}$. Training comprised batches of 4 pairs of volumes for up to 1500 epochs. Nevertheless, to prevent overfitting and to ensure the model's generalization, we used an early stopping strategy, causing the training to halt at the 410th epoch, corresponding to 1 day of training.

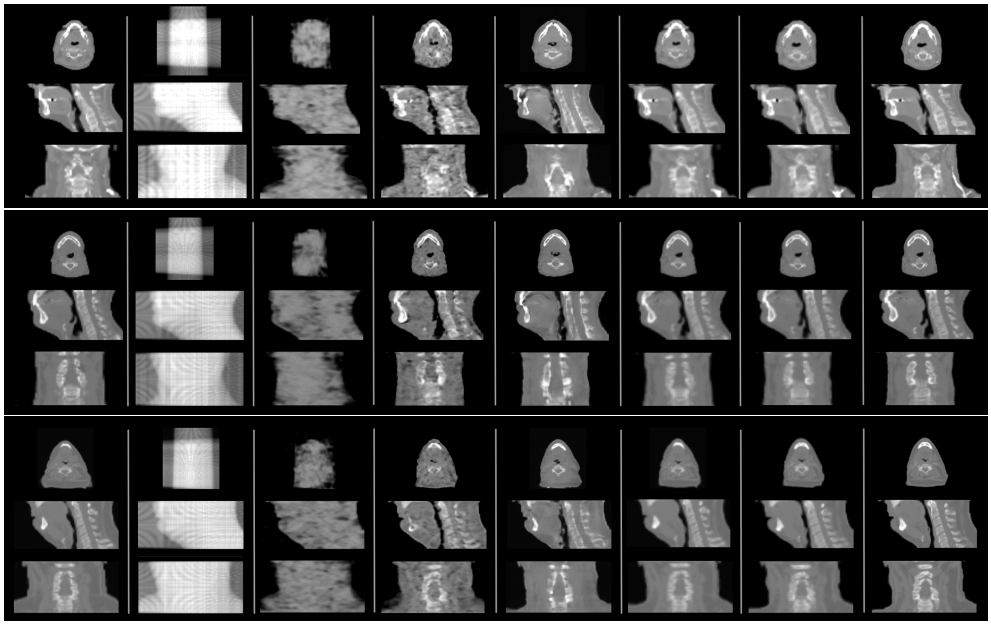

Figure 4.9: **Visual analysis of recovered volumes from two projections by previous methods and our approach.** In the absence of a pre-captured CT volume, NeRP struggles due to lack of constraints. When exploiting the pre-captured CT volume, NeRP still tends to introduce artifacts in an attempt to align with the projections and alters the anatomy without ensuring anatomical accuracy. In contrast to X2Vision, our method predicts a reconstruction that captures patient-specific details and nuances. 2D3DNR results in deformations that do not adequately match the anatomy.

Volume Recovery. For the volume recovery, we conducted the optimization process on a 16GB V100 GPU, utilizing the Adam optimizer with a learning rate of $1\mathrm{e}{-3}$. Through a grid search on an external validation set of 10 patients, we determined the optimal weights for our model: $\lambda_2 = 10$, $\lambda_p = 0.1$, $\lambda_w = 0.1$, $\lambda_c = 0.05$, $\lambda_n = 10$. These match the ones found for X2Vision.

For our volume recovery, we kept consistent weights for both the warm-up optimization, which excluded the pre-captured volume, and the main optimization, which incorporated it. Through a grid search on the validation set, we found that a 10-step warm-up was optimal. Without it, deformations often started from misaligned volumes, leading to suboptimal results. Conversely, an extended warm-up, like 50 steps, possibly trapped the optimization in local minima, yielding results too similar to a reconstruction with solely generative model followed by one deformation, not completely exploiting the joint optimization with concurrent priors. Our optimization runs for 100 steps, starting from the average latent vector.

Table 4.1: **Reconstruction metrics on volumes from two projections by previous methods and our approach.** Standard deviations are provided in parentheses. (w/) and (w/o) stand for the use or not of the pre-captured volume respectively.

Method	PSNR (dB) ↑	SSIM ↑
Backprojection	10.29 (±0.5)	0.23 (±0.01)
NeRP (w/o) [Shen, 2022a]	19.81 (±1.7)	0.21 (±0.03)
NeRP (w/) [Shen, 2022a]	25.32 (±1.6)	0.34 (±0.02)
X2Vision [Cafaro, 2023d]	27.80 (±1.4)	0.89 (±0.03)
2D3DNR [Dong, 2023]	29.07 (±1.6)	0.92 (±0.02)
Ours	**33.23** (±**0.62**)	**0.96** (±**0.01**)

Table 4.2: **Rigid and deformation metrics on volumes from two projections by previous methods and our approach.** Standard deviations are provided in parentheses. (w/) and (w/o) stand for the use or not of the pre-captured volume respectively.

Method	Dice ↑	
	Mouth	Larynx
2D3DNR [Dong, 2023]	0.91 (±0.03)	0.80 (±0.07)
Ours	**0.95** (±**0.01**)	**0.91** (±**0.02**)

Method	Rigid Registration Error (6 DoF)	
	Rotation (°) ↓	Translation (mm) ↓
2D3DNR [Dong, 2023]	0.52 (±0.29)	0.88 (±0.45)
X2Vision [Cafaro, 2023d]	0.45 (±0.31)	0.50 (±0.26)
Ours	**0.16** (±**0.15**)	**0.20** (±**0.07**)

4.5.3 Metrics

Metrics. We assessed the reconstruction performance using two quantitative metrics: PSNR, which quantifies reconstruction error, and SSIM, which measures the perceptual quality of the images. We also evaluated the accuracy of the deformation between the pre-acquired volume and the recovered volume for the two methods that estimate this deformation: 2D3DNR and ours. To this end, we consider the Dice score for the mouth and the larynx, two structures that are likely to deform significantly. To compute it, we segmented these structures on the groundtruth and recovered volumes using a trained U-Net model using about 1000 head-and-neck CTs.

Additionally, we compared 3D (6DoF) rigid registration differences, between initial full CT scans and our reconstructions against the ground truth.

This analysis highlights the precision of our method in capturing the nuanced critical anatomical features.

4.5.4 Results and Analysis

4.5.5 3D/3D Deformable Registration

For evaluating our 3D/3D deformable registration model, we calculated the mean squared error specifically within the designated region of interest. The mean squared error (MSE) was $6.7\mathrm{e}{-4}$ for the training dataset and $1\mathrm{e}{-3}$ for the validation dataset. Using the same group of 70 patients previously selected for evaluating our complete method of volume recovery, we attained in the targeted zone of interest excellent metrics: a PSNR of 37.64 (± 2.50) and a SSIM of 0.99 (± 0.01).

4.5.6 Volume Recovery

Tables 4.1 and 4.2 report the quantitative results. We detail below the methods we compare to and discuss their results after they were retrained on our data. Figure 4.9 compares visually our reconstruction to these methods on several examples. Additional results and reconstructions are provided in the supplementary material, but we summarize below our visual analysis of the results.

The **backprojection method** is a very simple baseline presented in 3.1.2. It estimates the value of each voxel as the average of the values at the projected voxel locations in the input X-ray projections. When enough input projections are available, this method can provide satisfying results. However, it fails when only two projections are used.

The **NeRP** method [Shen, 2022a] optimizes the 3D volume to match the projections. It also struggles when very few projections are given since they lack prior anatomical knowledge. Even when conditioned on the pre-captured volume, it is often not able to eliminate the many artefacts.

We also considered our previous **X2Vision** method to highlight the advantages of exploiting the pre-captured volume as we do—which X2Vision does not. It provides a reasonable reconstruction but still misses specific details and abnormalities.

2D3DNR [Dong, 2023] predicts in a feedforward way the deformation between the pre-captured 3D volume and the new one given the precaptured volume itself and the available biplanar projections. Since the original code was unavailable, we used the same VoxelMorph backbone as ours to reimplement the 3D/3D registration method. Further details can be found in the Appendix 4.7. The volumes predicted by 2D3DNR do not reproject well on the input projections in general and the predicted deformations can be inconsistent. Because it is a feedforward method, it also tends to generalize poorly. Our method recovers better the deformation of the tissues.

Like X2Vision and NeRP, our method optimizes on the volume during inference for consistency with the input projections, which helps generalization. It also introduces a prior on the anatomical volume thanks to its GAN, in a way related to X2Vision. Our method has however an original way to exploit the pre-acquired volume by controling its deformations. This contrast with 2D3DNR, which takes this volume as input to a

Table 4.3: **Ablation Study.** This table shows the contribution of our two priors. More details can be found in Section 4.5.9.

Metric	deformation of the pre-acquired volume without prior	generative model only	generative model followed by deformation	deformation of the pre-acquired volume only	full method
PSNR (dB) ↑	27.04 (±1.9)	27.80 (±1.4)	29.24 (±1.8)	30.75 (±1.19)	**33.23** (±**0.62**)
SSIM ↑	0.88 (±0.03)	0.88 (±0.03)	0.92 (±0.02)	0.93 (±0.01)	**0.96** (±**0.01**)

feedforward process, and with NeRP, which uses this volume only as conditioning. Our approach appears to be more powerful as it yields the best results.

Although X2Vision provides close reconstruction, it lacks details. Our method integrates patient-specific details to surpass the constraints of the generative model manifold, which might not capture the patient's unique anatomy or abnormalities accurately. By leveraging the pre-acquired volume, our method obtains a more accurate depiction of the patient's real anatomy rather than depending on a learned manifold. This focus on patient specificity is crucial for achieving detailed and realistic anatomy reconstructions.

In contrast to 2D3DNR, we adopt an optimization strategy to inform the deformation prior with the capabilities of the generative model. This model lays down a realistic 3D support for the optimization process, ensuring the deformations are not just plausible but supported by the anatomical frame of the generative model. This leads to more precise reconstructions, showcasing a significant step forward in anatomical fidelity.

4.5.7 Validation for Medical Applications

As shown in Table 4.2, our method achieves precise rigid registration with average errors well below 1mm and demonstrates strong deformation accuracy, evidenced by high Dice coefficients and implicit accurate segmentation of critical organs at risk. This high level of detail and correspondance highlights our approach's capability to capture complex patient anatomy.

By achieving high-quality 3D reconstruction from just biplanar projections, augmented with pre-captured planning CT data, we can enable accurate 3D rigid registration and facilitate daily monitoring of patient changes without the need for full 3D acquisitions. This method paves the way for adaptive radiotherapy, allowing for potential replanning based on reconstructed 3D volumes.

The degree of alignment achieved may not only support dose accumulation and trigger the need for adaptation and monitoring but also allow for direct replanning based on well-segmented and accurately deformed structures.

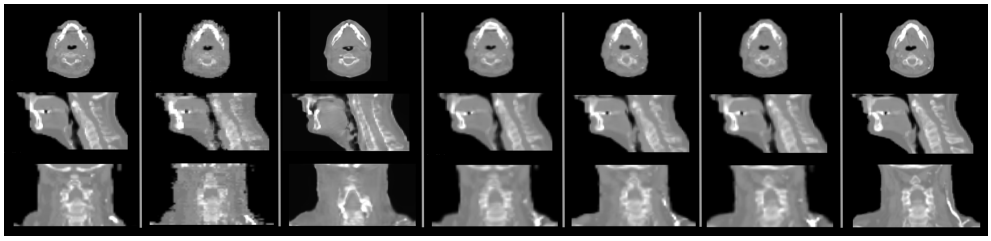

Figure 4.10: **Visual Analysis of the Ablation Study.** Deforming the pre-acquired volume without any prior results in erratic and anatomically inconsistent changes. Reconstruction solely with the generative model may overlook details and lead to mismatches. Deforming the pre-acquired volume on it introduces patient-specific features but may retain initial misalignments. While introducing prior on deformation aids guiding the direction process, it leads to unnatural distortions of body contour and bone structures. Our method, by leveraging both anatomical and deformation priors, yields more realistic and anatomically preserving results.

Table 4.4: **Inference time** for the different methods.

Method	Inference Time
Backprojection	<0.5 sec
NeRP (w/o prior volume)	4 min
NeRP (w/ prior volume)	7 min
X2Vision	30 sec
2D3DNR	<0.5 sec
Ours	1 min

4.5.8 Inference Time

Table 4.4 provides a comparison of inference time for the different methods. Our method recovers high-quality volumes in only 1 minute. While some other methods are faster, the trade-off fidelity/runtime is well acceptable as clinical CBCT acquisition and FDK reconstruction [Feldkamp, 1984a] currently requires more than 2 minutes, as previously mentioned.

4.5.9 Ablation Study

Table 4.3 presents an ablation study highlighting the benefits of our loss function in Eq. (3.35) by comparing it to different possible variants. Figure 4.10 presents a visual comparison of the results obtained with these variants. We considered four variants. Implementation details can be found in the Appendix. The reader should compare the loss functions for these variants to the loss function we introduced in Eq. (3.35):

- 'deformation of the pre-acquired volume without any prior': This variant returns volume $\mathcal{W}(v^-, \phi^*)$ with $\phi^* = \text{argmin}_\phi \sum_i \mathcal{L}_i(\mathcal{W}(v^-, \phi), I_i)$, where $\mathcal{W}(v^-, \phi)$ applies a deformation field ϕ to the pre-acquired volume v^-. This approach retrieves the deformation parameters iteratively without any prior learning. The retrieved deformations tend to be erratic, blending structures and leading to artifacts that compromise anatomical accuracy.
- 'generative model only': This variant returns $v(\mathbf{g}^G)$ with $\mathbf{g}^G = \text{argmin}_\mathbf{g} \sum_i \mathcal{L}_i(v(\mathbf{g}), I_i) + \mathcal{R}(\mathbf{g})$. It uses only the generative model to predict the volume and corresponds to the X2Vision method.
- 'generative model followed by 1 deformation': This variant returns volume $S(v^-, v(\mathbf{g}^G))$. This approach deforms the pre-acquired volume to fit the generative model's reconstruction, introducing patient-specific details but potentially retaining initial mismatches. This shows the advantage of combining volume v^- and the generative model $v(\mathbf{g})$ during optimization.
- 'deformation of the pre-acquired volume only': This variant returns volume $S(v^-, \mathbf{v}^*)$ with $\mathbf{v}^* = \text{argmin}_\mathbf{v} \sum_i \mathcal{L}_i(S(v^-, \mathbf{v}), I_i)$, where $\mathbf{v}$ is a volume represented by a voxel grid, with each voxel encompassing an intensity to optimize. This approach uses only the pre-acquired volume and the spatial transformer, but not the generative model. This results in local deformations that are not anatomically realistic, such as bone extensions or body contour distortions, stemming from its lack of anatomical prior.

The quantitative results in Table 4.3 clearly show that our loss function exploits both priors well.

4.6 Conclusion and Discussion

By managing to exploit both patient-specific data and anatomical constraints, we achieve unmatched accuracy in anatomical reconstructions, potentially avoiding the need for intensive 3D scans. Our method promises improved patient care with daily adjustable treatments for adaptive radiotherapy, enhancing precision and outcomes while reducing treatment times and radiation exposure.

The core of our method lies in the concurrent generation and deformation of anatomy, achieved through an iterative, fully unsupervised process that ensures consistency with the original projections. Priors play a crucial role: anatomical guides shape the deformation, and deformation, in turn, guides the anatomy to converge to a close solution. By inte-

grating patient-specific details, we preserve critical anatomical nuances, resulting in more accurate reconstructions and better overall outcomes. This pipeline has proven to be robust, demonstrating excellent results and impressive alignment with real anatomical structures.

Extension to Clinics Our results are promising and indicate potential for clinical application; however, further evaluation with real patient data and additional clinical metrics is necessary. Our method demonstrates the capability to replace extensive CBCT scans by reconstructing 3D images from biplanar X-rays, initially for rigid positioning. For full integration into adaptive treatment, dosimetry considerations are crucial. The improved accuracy of our approach compared to X2Vision suggests it could provide more precise dosimetric analysis for accurate dose accumulation and for triggering the need for adaptation.

Furthermore, by directly deforming the planning CT, our method facilitates structure propagation, potentially enabling direct replanning based on the reconstructed images. Additional evaluation is required to accurately assess tumor deformation. While relying solely on surrounding structures may introduce some uncertainty in tracking deformation, this approach could still effectively guide adaptation or replanning. Accurate structure propagation could enable adaptive radiotherapy similar to CBCT, as discussed in Chapter 2.

To implement this approach in clinical settings, it's essential to adapt this method for real X-rays and address challenges such as calibration, scatter effects, noise, and variable patient positioning in biplanar systems. An adaptation will be introduced in the final chapter 5, where we propose validating our technique using real biplanar X-rays from the Exactrac system.

Temporal Effects Our method is flexible regarding the pre-acquired volume used; it does not need to be the initial planning CT scan. It could be a volume from the previous day, such as a prior reconstruction or CBCT, which would limit the degree of transformation required and potentially improve results. The method has been evaluated with significant anatomical changes occurring after several weeks of treatment, encompassing a wide range of transformations. In typical clinical scenarios with smaller, incremental changes during treatment, performance could be even better. Evaluating reconstruction quality based on treatment progression or the difference between the pre-acquired volume and the one being reconstructed could further assess the potential of our method for clinical application.

Regularization with Priors on Anatomy and Deformation In this inherently ill-posed task, regularization has proven to be essential. By incorporating priors on both anatomical structures and possible deformations, we significantly reduce the degrees of freedom, allowing the generative and deformation models to better capture realistic anatomy and potential variations. Improving these models to better represent the 3D manifold and possible deformations would further refine the quality of the results.

VoxelMorph has demonstrated strong capability in learning intra-patient deformations, showing near-perfect results on the test dataset for 3D-to-3D deformations. Adding more regularization could help better model different types of deformations, and incorporating more data with diverse transformations would further enhance its performance. Anatomical guidance, such as organ masks, could also improve training by adding constraints, similar to how VoxelMorph can incorporate masks. Hybrid models that integrate biomechanical principles, along with more advanced deformation networks (e.g. with transformers), could push performance even further.

Despite the problem being ill-posed and non-convex, relying on the learned manifolds of 3D structures and deformations significantly reduces the solution space. A study on uncertainty could help identify the range of potential solutions when priors are introduced, possibly through variational inference, as potentially done with X2Vision. Despite regularization, regions like internal tissues may still have zones of uncertainty, especially with only two X-rays providing limited supervision.

The effectiveness of anatomical regularization depends on how well the generative model learns fine details and internal distributions. If it successfully captures basic structures like bone, tissue, and air, regularization can be applied at that level. However, extending the model to learn more complex internal distributions—such as muscles, tumors, and cartilage—allows for even finer regularization. More detailed learning leads to better results overall. The level of detail learned by the GAN directly influences the precision of deformation models. In complex internal structures, improving anatomical precision requires more learning and finer regularization, extending beyond broad structures to finer details.

As mentioned earlier, methods like 2D3DNR generate pseudo-3D structures, but they are not necessarily realistic. They serve as methods to leverage 2D information into 3D, but unsupervised training using a true generative model, like ours, ensures realistic 3D outputs and adds valuable constraints—going beyond just 2D-to-3D conversion. Following this idea, another approach similar to ours and 2D3DNR is DiffRecon [Sun, 2024a], which uses a reconstruction-registration model employing diffusion for more realistic generation, guiding deformations for CT reconstruction from a few planar DRRs.

Learning Biases with Population-Based Priors As with other learning-based approaches, relying on priors introduces challenges with out-of-distribution cases, leading to biases in specific population groups or types of deformations. Compared to methods like X2Vision, our approach is more robust thanks to the introduction of pre-acquired CT data, which imposes stronger constraints. Learning deformations is also easier and less prone to out-of-distribution issues, or at least the effects are more subtle. These constraints guide the model more effectively, reducing the risk of failure seen in other methods. While some unrealistic deformations in tissues may still occur, the results are generally more consistent, as shown by improved metrics.

However, not all possible deformations or anatomies are captured, and this can result in coarse reconstructions for outliers or rare conditions. Biases related to factors such as gender or race may contribute to these errors. Addressing this requires larger and more diverse datasets, or incorporating specific priors for abnormalities. As learning improves, so will the quality of manifold approximations, but additional validation on broader datasets is crucial to better represent the full population.

Deformable Registration Limitations Relying on deformable registration has limitations, especially with large transformations such as significant weight loss, tissue reduction, or artifact/metal removal. These cases go beyond simple deformations and involve non-diffeomorphic changes that deformation models struggle with due to one-to-many mappings or loss of matter. Direct signal reconstruction methods would likely yield better results in these scenarios. Optimal transport methods [Cuturi, 2013] could be more effective for handling such transformations.

However, with only two projections, deformations already provide a very useful approximation, offering anatomical insights compared to reconstruction alone. An alternative idea is to have the generation guided by deformation, rather than deformation guided by generation. Instead of projecting the deformed CT onto the GAN, the GAN could generate a structure close to the CT using deformation as a guide while still matching the projections. This would help incorporate patient-specific features and ensure non-deformable regions are accurately reconstructed. This could balance the generation with real anatomical details while optimizing projections, offering a stronger approach by combining both techniques.

This approach depends on the model's ability to learn the distribution with fine details. Deformation contributes patient-specific information that the GAN might miss if its representation is inaccurate, especially in out-of-distribution cases or if the learning model isn't complex enough. By relying directly on the learned manifold, the generation may tend to match the deformed CT and the actual projections, but it remains constrained by the manifold and could miss important patient-specific details. In contrast, deforming the CT onto the generation, as we do here, guides the generation closer to the patient, projected onto the manifold. This approach introduces true patient-specific details and significantly improves robustness.

Another approach could involve direct reconstruction by integrating CT data with projection embeddings for 3D reconstruction. This might use a 3D conditional GAN for 2D-to-3D conversion, followed by fusing pre-captured data with pseudo-3D reconstructions to improve accuracy. Techniques like diffusion models, such as ControlNet [Zhang, 2023c], could further enhance this process by guiding the generation with both CT and projection embeddings. However, deformable registration is typically preferred, as it makes the reconstruction problem much more well-posed.

4.7 Appendix

4.7.1 Implementation Details for Baselines and Ablation Study

Baselines. The backprojection method was designed using the backprojector inspired by [Peng, 2021]. Given the unavailability of a 2D3DNR implementation, we undertook its reimplementation. This included the embedding of the projections into a 3D feature map and substituting the original spatial transformer with the Voxelmorph architecture we used for our model. In this setup, 3D embedding from projections serves as a substitute for high-quality CT in the reconstruction area. The loss function employed is identical to the one used in our method. For the training and evaluation of this feedforward method, we crafted paired projection-reconstruction sets by generating projections from our extensive longitudinal data.

Methods in the Ablation Study.

- 'deformation of the pre-acquired volume without any prior': This version entails optimizing the deformation field to align the pre-acquired volume with the projections. The process begins with the field initially set to zeros, indicating the absence of any deformation. We optimize this field using gradient descent with Adam at a learning rate of $1\mathrm{e}{-3}$. The process involves 2000 steps, successfully achieving convergence within a time frame of approximately 115 seconds.

- 'generative model only': This version use the same values for the loss weights as those used for the X2Vision method.

- 'generative model followed by 1 deformation': This version performs a reconstruction using the generative model. Subsequently, it deploys our 3D/3D deformable registration model to align the pre-captured volume with it.

- 'deformation of the pre-acquired volume only': This version optimizes a 3D volume to ensure the pre-captured volume deforms and aligns with the projections. We initialize the volume to zeros, which correspond to the intensity of the tissue (within a range [-1,1]). We optimize this volume using gradient descent with Adam at a learning rate of $1\mathrm{e}{-1}$. The process involves 100 steps, successfully achieving convergence within a time frame of approximately 95 seconds.

4.7.2 Additional Visual Results

Figure 4.11: **Additional visual analysis of recovered volumes from two projections by previous methods and our approach.** Our recovery results consistently exhibit closer proximity to the ground truth, featuring enhanced alignment of structures, improved detail fidelity, and no artifacts.

Chapter 5

Towards Real-World Clinical Translation

This chapter explores how we adapt our 3D reconstruction frameworks, X2Vision and XSynthMorph, for real clinical workflows in adaptive radiotherapy using biplanar X-rays. Initially developed with DRRs in controlled conditions—fixed poses, full field of view, and a defined region of interest—it demonstrated the feasibility of tackling the highly ill-posed 3D reconstruction problem by focusing on main complexities.

However, translating these methods to real biplanar X-ray systems is crucial for clinical application, but it brings new challenges. Real X-rays feature partial fields of view, varying regions of interest, non-coplanar imaging angles, and added noise and calibration differences—all of which add ambiguity to an already under-constrained task. Our goal is to expand our reconstruction approach to address these challenges, enabling seamless integration with existing biplanar radiotherapy systems.

To bridge the gap between controlled and real-world conditions, we create DRRs aligned with real X-rays through geometric matching and a domain translation network. We further enhance robustness by adapting our generative and deformation models and incorporating pre-positioning adjustments. This chapter centers on XSynthMorph, as it delivers the most robust performance under these conditions.

Here, we outline our approach to replicating real biplanar systems and adapting our frameworks for clinical application We discuss the challenges, innovations, and experimental outcomes involved in integrating these methods into clinical radiotherapy workflows.

Contents

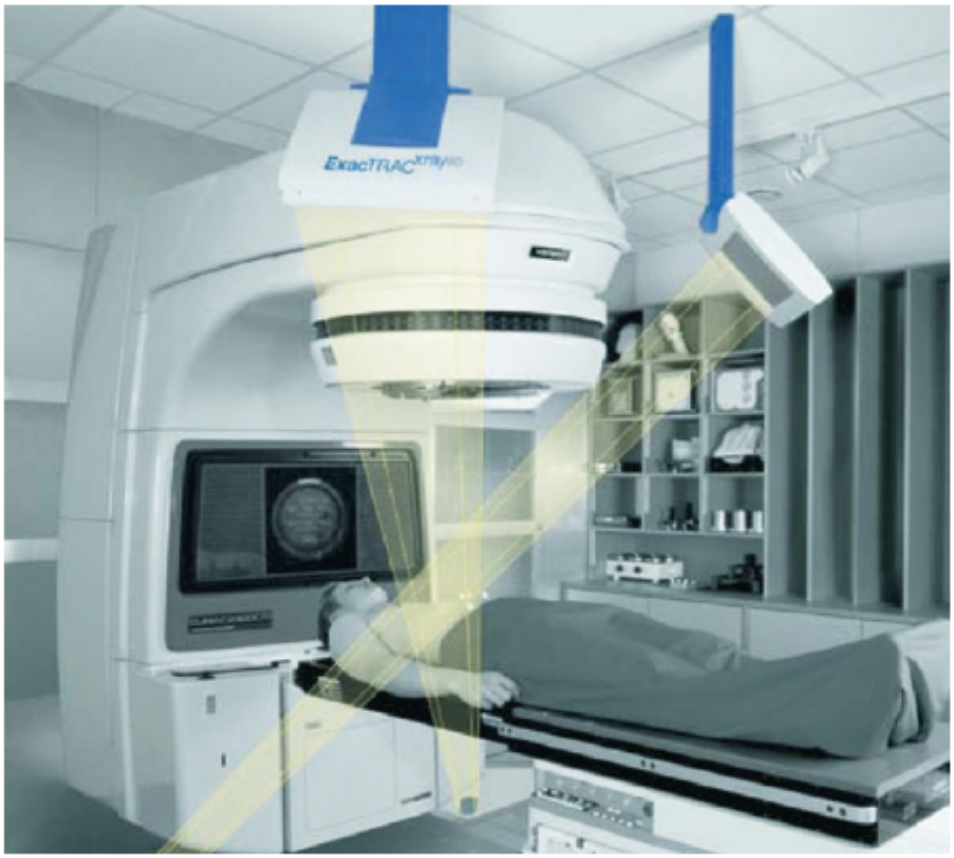

Figure 5.1: The Linac with the ExacTrac system, featuring two oblique, non-coplanar X-ray tubes that capture bone-focused images of the patient's anatomy in a limited field of view.[Jin, 2008]

5.1 Match Real Biplanar X-Rays

5.1.1 Presentation of the ExacTrac Biplanar System

The Brainlab ExacTrac [AG, 2024], introduced in Chapter 2 is a widely utilized IGRT system designed for accurate patient positioning during radiotherapy. Jin et al. [Jin, 2008] well describes and analyses this system. This system integrates two primary components:

the infrared-based (IR) optical positioning system and the biplanar non-coplanar X-ray imaging system.

Figure 5.1 shows a Linac with the ExacTrac system, showing the oblique configuration of the X-ray imaging devices.

The IR optical positioning system consists of two ceiling-mounted infrared cameras that track reflective markers placed on the patient's skin, along with a reference star attached to the treatment couch. These markers allow for the precise initial setup of the patient and real-time monitoring during treatment. The IR system provides high spatial resolution (better than 0.3 mm), which ensures stability in patient positioning even during movement, such as respiratory motion, making it particularly useful in gated radiotherapy.

The X-ray imaging system features two floor-mounted, obliquely positioned, non-coplanar X-ray tubes that capture radiographic images focused on the patient's bony anatomy. These images have a limited field of view, approximately 10 cm at the isocenter, specifically targeting the region near the tumor and aimed to align with the PTV's isocenter.

After the initial coarse setup using IR markers, the positioning is refined with the use of X-rays. This X-ray fusion-guided adjustment is classically performed once per treatment session, just before radiotherapy begins, but can be repeated during the fraction.

The system's main objective is to align the radiotherapy target at treatment time as closely as possible with the planned position in the treatment machine's coordinate system, specifically aligning the PTV isocenter with the imaging system's isocenter. Captured radiographs are then fused with pre-existing CT simulation images using either 3D or 6D fusion algorithms. The 3D method corrects for translational shifts only, while the 6D method also adjusts for rotational deviations, including pitch, yaw, and roll [Jin, 2008].

This alignment is achieved through iterative optimization of six parameters (three translations and three rotations) to minimize the cross-correlation between DRRs from the CT and actual X-ray images [Lemieux, 1994], focusing on aligning the patient's bony structures for precise 3D registration. The fusion matrix transforms the CT data to the actual X-ray image, which likely shows the patient in a misaligned position. The inverse of this fusion result is then applied to the treatment couch, bringing the patient into the correct alignment for treatment.

Figure 5.2 shows an example of a DRR at the CT isocenter before and after applying 6D correction, aligned with the real X-rays.

The 6D fusion method offers a significant improvement in localization accuracy over the 3D method by accounting for both rotational and translational deviations. Studies demonstrate that 6D fusion can achieve sub-millimeter accuracy. For example, in phantom studies, localization accuracy was within 1 mm, even when rotational deviations were introduced. Clinical studies [Jin, 2006] in patients with cranial and spinal lesions further validate this, showing that 6D fusion can correct rotational deviations of up to 4° more effectively than 3D fusion. This is especially critical when large rotational shifts occur during patient setup.

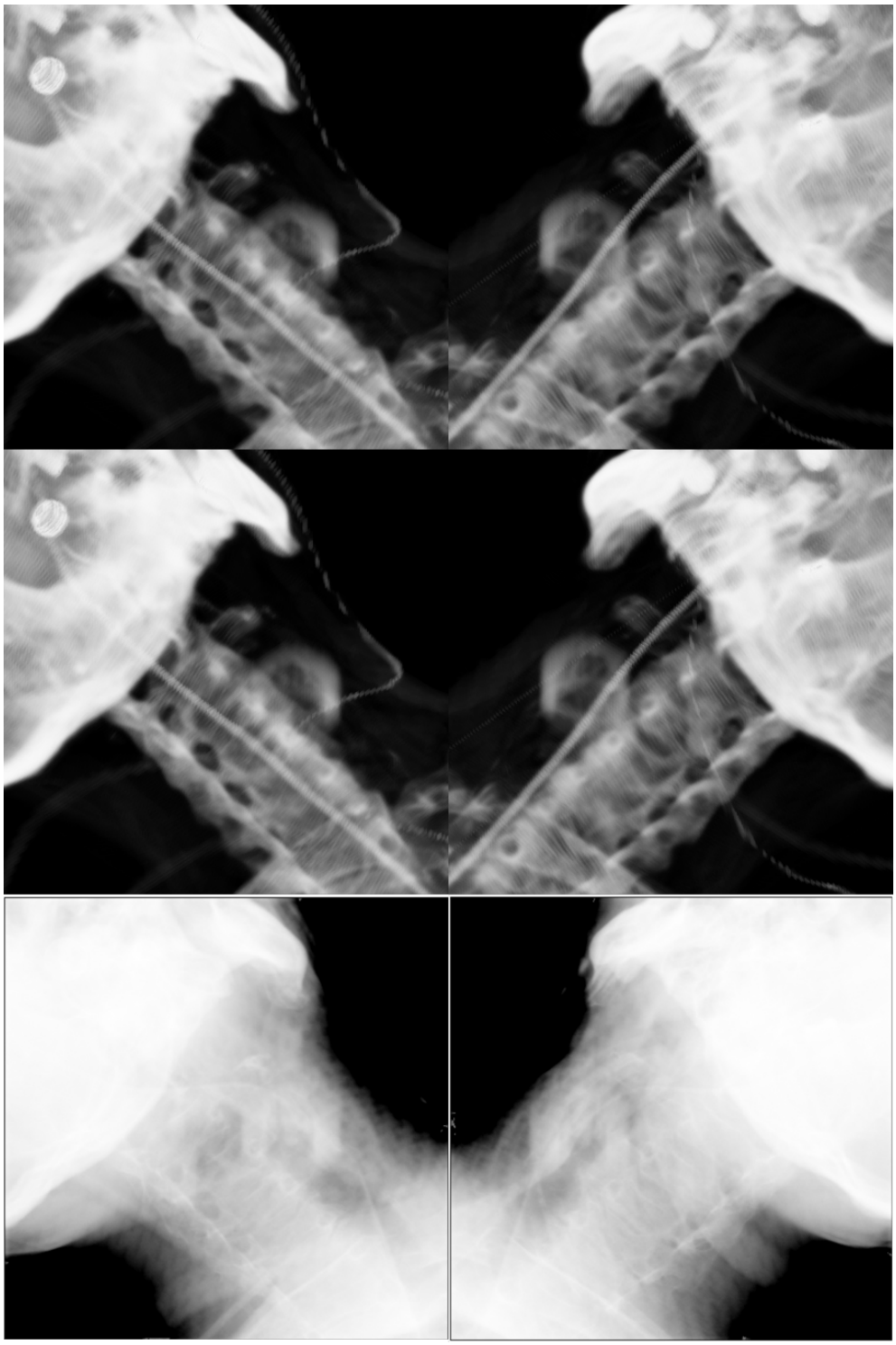

Figure 5.2: DRRs before registration (top), DRRs after 6D table correction (middle), both focusing on bones, and real initial X-rays acquired with Exactrac (bottom).

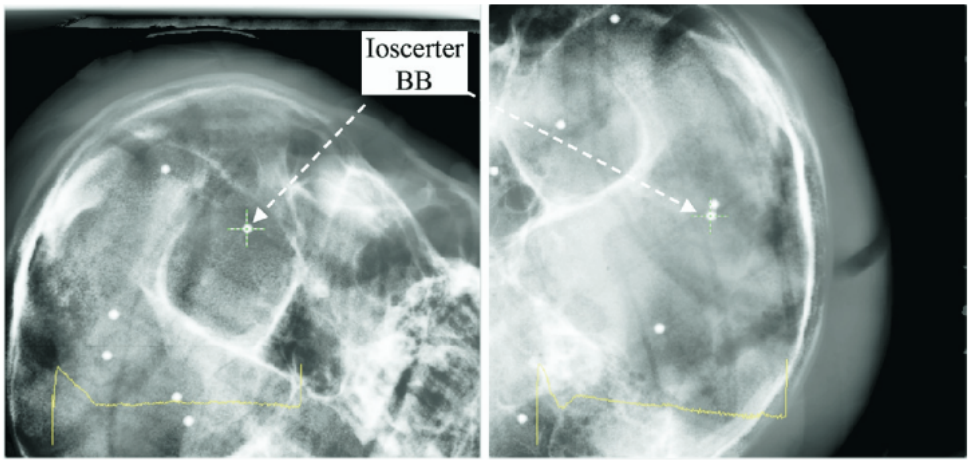

Figure 5.3: Biplanar verification X-ray images evaluating localization accuracy by measuring the distance between the X-ray system's isocenter (represented by the cross) and the treatment isocenter (center of the BB) [Jin, 2008].

To verify alignment, the system can use a small metal ball bearing (BB) placed at the treatment isocenter as a fiducial marker. The BB acts as a precise reference point for aligning both the X-ray and treatment isocenters. Figure 5.3 illustrates this alignment, showing how the X-ray system's isocenter (marked by the cross) closely aligns with the treatment isocenter (center of the BB) [Jin, 2008].

The ExacTrac X-Ray 6D system is ideal for treating targets connected to rigid bony structures, such as cranial and spinal lesions. However, image quality can be reduced in larger patients due to overlapping anatomical structures and the oblique positioning of the X-ray devices, with longer X-ray paths through the body complicating registration further. Additionally, when the tumor is not close to bony structures, soft tissue registration becomes less precise. Implantable markers can improve accuracy, but 3D tomographic imaging, like CBCT, is required to visualize internal anatomy accurately for precise soft tissue registration. For head and neck cases, a CBCT scan is often acquired after the initial Exactrac prepositioning to obtain a 3D estimate and refine the alignment.

5.1.2 Project like Real Biplanar Systems

Real Geometry

Our methods, X2Vision and XSynthMorph, were developed using a 90° face profile with a full field of view centered on the region of interest for reconstruction. However, the field of view and geometry of the ExacTrac system differ. To adapt our methods to real biplanar systems like ExacTrac, we need to generate DRRs that closely replicate the geometry, field of view, and energy of real X-rays. This approach will allow us to minimize the discrepancy between DRRs and real X-rays, enhancing the accuracy of our reconstructions. Our objective, therefore, is to produce DRRs that emulate those of the

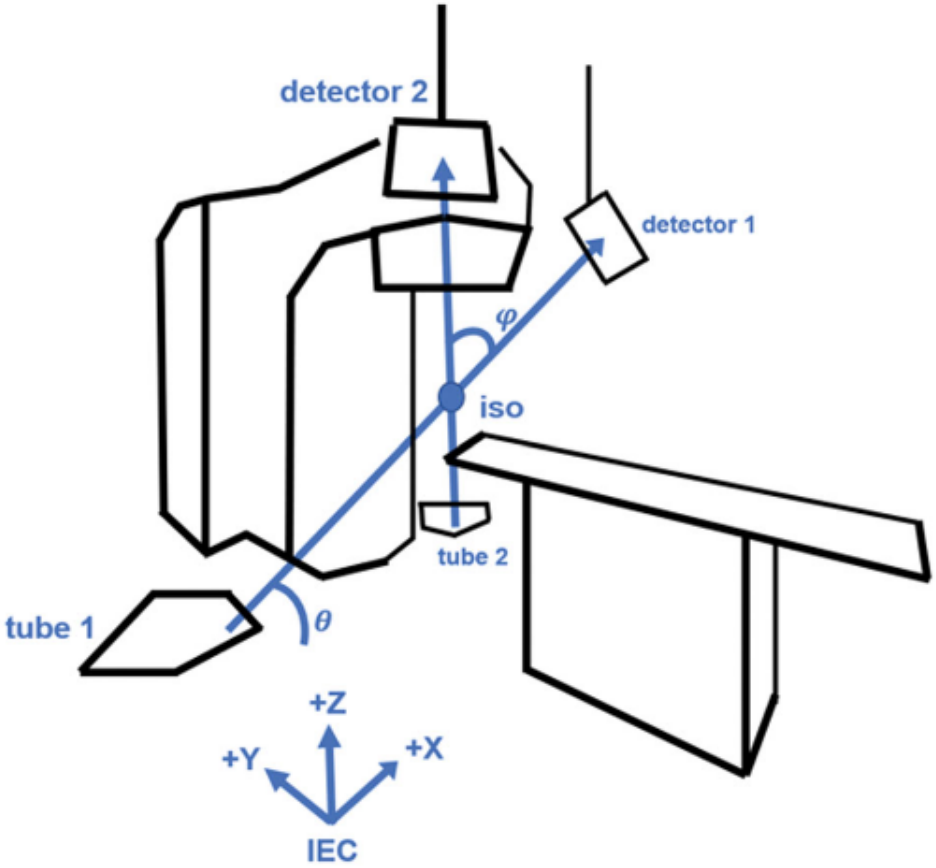

Figure 5.4: Exactrac system geometry. X-ray tube 1 is paired with flat panel detector 1, and X-ray tube 2 with detector 2. The central beamlines from both tubes intersect at the isocenter at a crossing angle ϕ. The oblique plane formed by these central beamlines has an incline angle θ relative to the floor [Charters, 2022]. In practice, θ is set to 53° and ϕ to 63°, both well below 90°. This configuration adds ambiguity to the reconstruction.

ExacTrac system for any patient position.

Charters et al. [Charters, 2022] developed a method to partially address this by creating a DRR projector that accurately replicates ExacTrac DRRs, specifically focusing on bone structures to match those used for rigid registration in real X-ray systems. Their projector generates precise DRRs by using either in-room measurements or exact configuration values.

As shown in Figure 5.4 the ExacTrac geometry relies on key parameters like the source-to-image distance (SID), source-to-object distance (SOD), oblique plane angle (θ), and central beamline crossing angle (ϕ). With these measurements, we can generate stereoscopic DRRs for a given CT volume and isocenter position. Due to the system's symmetry, focusing on one tube-detector pair is often sufficient. Here, SID is the distance from the X-ray source to the detector center, while SOD is the distance from the X-ray

source to the isocenter. Each X-ray beam originates from a point in the X-ray tube and strikes the flat-panel detector. The central beamline is defined as the ray that hits the center of the detector. Unlike traditional setups with orthogonal, face-profile angles as used in our initial work, ExacTrac employs oblique, non-coplanar angles which allow the rotation of the Linac. In practice, θ is set to 53° and ϕ to 63°, both well below 90°. This configuration adds ambiguity to the reconstruction process, as it crosses more tissues or reduces the disentanglement of depth information.

While the ExacTrac geometry can be modeled using first principles, system geometry varies between treatment centers, requiring adaptation for each local setup.

Key parameters for DRR generation are stored in ExacTrac configuration files under calibration matrices produced for each session. By extracting calibration or renderer matrices detailing the flat panel and X-ray tube geometry, we can replicate the projection geometry in our custom DRR generators.

A renderer matrix M [Charters, 2022] integrates essential projection, translation, rotation, and scaling matrices, covering both extrinsic and intrinsic parameters.

Extrinsic geometry refers to the parameters that define the position and orientation of the X-ray source and detector relative to the object (in this case, the patient or CT volume). This includes the translation and rotation matrices, which transform the object from its world or patient-centered coordinate system to the coordinate system of the imaging setup (e.g., the X-ray detector). Intrinsic geometry, on the other hand, represents the internal parameters of the imaging system, such as the scaling and projection matrices, which define how the object coordinates are projected onto the detector plane.

The renderer matrix M transforms coordinates from the CT volume to the detector coordinates, ensuring that the digitally reconstructed radiographs (DRRs) align accurately with the geometry of actual X-ray images. The general form of this matrix is [Charters, 2022]:

$$M = S_{NDC} S_{det} P R T, \tag{5.1}$$

where:

- T (Translation Matrix): This matrix accounts for shifts between the CT scanner and the isocenter of the X-ray imaging system. It translates the coordinates to ensure correct alignment with the imaging system.
- R (Rotation Matrix): This matrix is based on the direction cosines of the detector, defining the orientation of the detector relative to the patient. It adjusts the coordinates to account for any rotation of the imaging system.
- P (Projection Matrix): This matrix simulates how the X-ray source projects the 3D CT data onto the 2D detector plane. It is essential for accurately capturing the geometry of the X-ray projection and ensuring the spatial relationships in the image are preserved.

- S_{det} (Scaling Matrix for Detector): This matrix converts coordinates from the intrinsic resolution of the detector to pixel spacing. It ensures that the DRRs are appropriately scaled to match the physical dimensions of the detector.

- S_{NDC} (Scaling Matrix for Normalized Device Coordinates): This matrix converts the detector coordinates to normalized device coordinates (NDC), which typically range from -1 to 1. This standardization is crucial for consistent image output across different detectors and resolutions.

By decomposing the ExacTrac renderer matrices, we can extract the precise geometric parameters required for generating accurate DRRs [Charters, 2022]. One effective method for achieving this decomposition is RQ decomposition, which separates the matrix into an upper triangular matrix R, and an orthogonal matrix Q. This decomposition isolates the parameters that define how the X-ray source projects onto the detector and how the CT volume is translated and rotated relative to the imaging system. We can express the projection matrix M as:

$$M = R \cdot Q.$$

We could also use the more precise Perspective-n-Point (PnP) methods [Lepetit, 2009], widely used in computer vision. By leveraging the projection matrix and correspondences between 3D points and their 2D projections, PnP techniques can accurately estimate the position of the X-ray sources relative to the detector and patient.

We tested both methods and found that RQ decomposition aligns with the approach used by Brainlab, making it the most reliable option.

The detector-scaled projection matrix is given by [Charters, 2022]:

$$P_{det} = S_{det} P = \begin{pmatrix} SID/s_x & 0 & c_x & 0 \\ 0 & SID/s_y & c_y & 0 \\ 0 & 0 & SID & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

This matrix corresponds to the projection matrix R in the RQ decomposition. The SID, represented by $SID = P_{det}(1,1)/s_x$, and the pixel shifts c_x and c_y are extracted directly from this matrix.

The focal points (source positions) can be derived from the full localization matrix, which combines the rotation (R) and translation (T) matrices. The localization matrix is given by:

$$L_{1,2} = \begin{pmatrix} p_x(X) & p_x(Y) & p_x(Z) & -t_{1,2} \cdot p_x \\ p_y(X) & p_y(Y) & p_y(Z) & -t_{1,2} \cdot p_y \\ d(X) & d(Y) & d(Z) & -t_{1,2} \cdot d \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

The 3x3 submatrix of $L_{1,2}$ corresponds to the rotation matrix Q from the RQ decomposition. The translation vectors $t_{1,2}$ (focal points) are derived by solving a system of linear equations and are then converted from the International Electrotechnical Commission (IEC) coordinate system to the head-first supine (HFS) coordinate system. The direction cosines matrices $D_{1,2}$ are obtained by transposing the rotation matrix Q [Charters, 2022].

By retrieving these parameters, we can accurately reconstruct the X-ray projection geometry, allowing us to generate DRRs that closely replicate the geometry of ExacTrac clinical X-ray systems.

To validate this approach, Charters et al. compared the generated DRRs with ExacTrac reference DRRs, showing excellent alignment and confirming the method's accuracy.

Generation of Realistic DRRs

The primary goal was to accurately replicate ExacTrac geometry. However, ExacTrac's current DRR rendering is simplified, utilizing a monoenergetic beam centered on bone. While this approach provides adequate precision for rigid registration, it does not fully capture real-world physics. To enhance alignment with actual X-rays, we developed more realistic simulations. With support from Brainlab and guidance from Charters et al., we refined our projection setup and incorporated key parameters into our more realistic, differentiable projector introduced in Section 3.4.3.

To improve realism, we additionally segmented air and implants (e.g., titanium dental implants) in the CTs and computed their absorption based on the correct energy levels of the spectrum, enabling accurate absorption estimation. Segmentation was achieved through thresholds on electron density (ED) units for more precision using calibrations curves from the centers.

For energy, ExacTrac's kV settings range from 60 to 120 kV, with mAs adjustable from 5 to 40 based on the indication and patient size. For our study, we used ExacTrac images at 100 keV and 10 mAs. Without direct access to the exact X-ray spectrum and filters, we simulated a 100 keV spectrum using SpekCalc [Poludniowski, 2009], with 5 keV binning to approximate a multi-energy X-ray source.

We generated DRRs centered on the isocenter with identical geometric parameters to those in the positioning plan. Since ExacTrac DRRs focus primarily on bones, we replicated this by producing DRRs excluding soft tissues for bone-centered comparison. Additionally, we generated DRRs that included soft tissues to further improve the match with real X-rays. An initial set of DRRs before 6D corrections is shown in Figure 5.5.

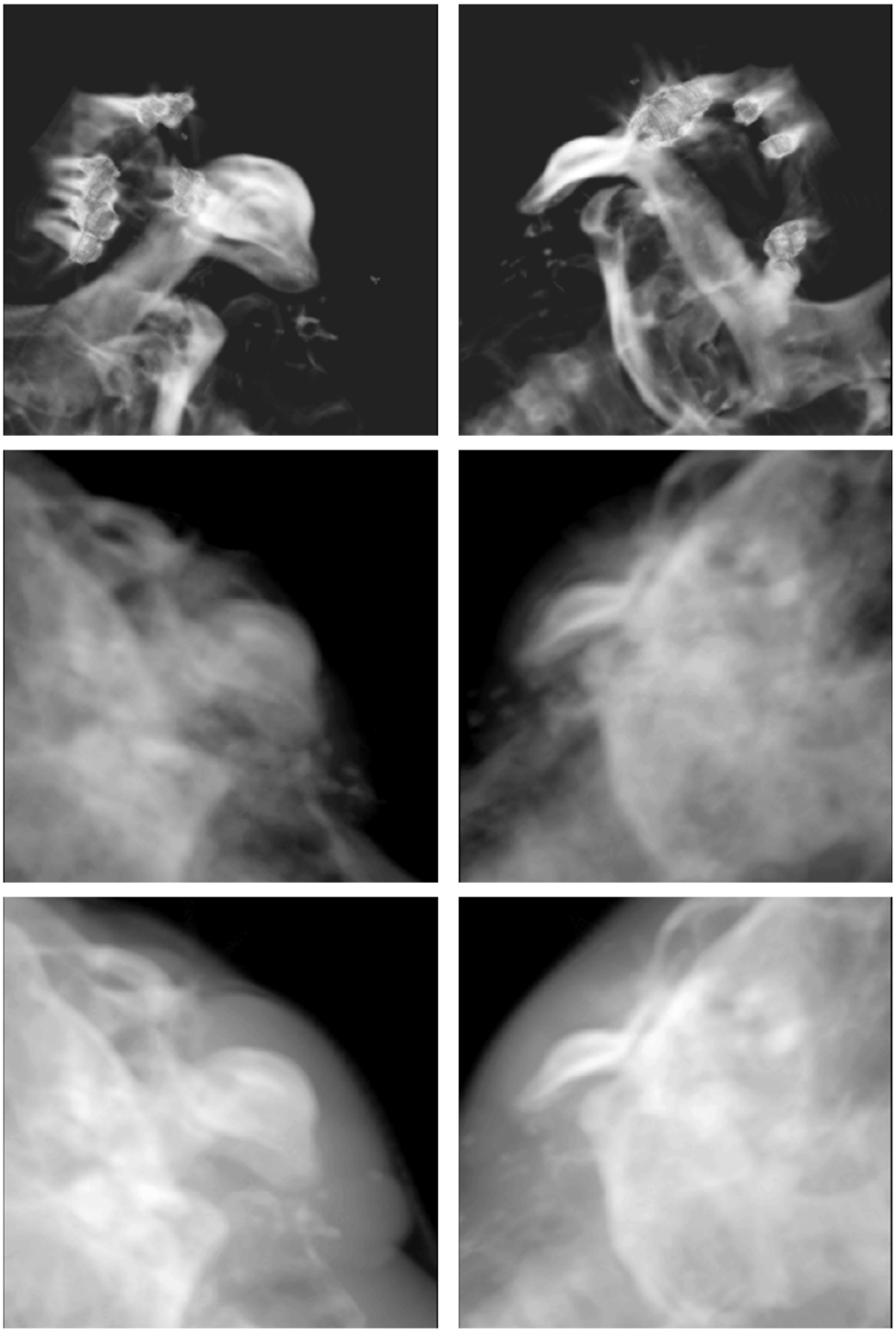

Figure 5.5: Comparison of our DRRs, without (middle) and with (bottom) tissue, with ExacTrac DRRs (top) before applying rigid 6D corrections.

ExacTrac's DRRs may use limited masks, omitting certain cartilage and bones. Our approach achieved an excellent match with ExacTrac's DRRs focused on bony structures.

Rigid and Deformable Alignment of DRRs with X-rays

To accurately generate DRRs that match real X-rays, we must align the DRRs to the patient's anatomy as shown in the X-rays. However, DRRs are generated from the planning CT's isocenter, while X-rays reflect the patient's position on the treatment table. This positional difference requires a 6D rigid correction to align the CT with the X-rays. Also, due to potential anatomical changes, a rigid alignment alone may be insufficient.

To address this, we obtain an anatomy of the day by acquiring a CBCT. Because CBCT images often have limited HU accuracy and artifacts affecting DRR quality, we deform the planning CT to match the CBCT, creating an updated virtual CT of the day aligned with the X-rays. By applying the initial 6D rigid correction to this virtual CT we get an aligned 3D anatomy with the X-rays, enabling the generation of DRRs that reflect the patient's current anatomy.

Figure 5.6 shows DRRs generated using this process, compared with real X-rays. While ExacTrac DRRs align well with bony structures after correction, our DRRs achieve improved alignment with both the bony and soft tissue structures, reflecting the patient's updated anatomy.

5.1.3 Domain Translation between DRRs and X-Rays

Discrepancies between DRRs and X-Rays

After aligning DRRs and X-rays both rigidly and elastically, achieving an even closer match is crucial for accurate 3D reconstruction. While robust loss functions like NCC allow to align bony structures well in 2D/3D registration, they fall short for precise 3D reconstruction. Small attenuation differences between DRRs and real X-rays can lead to substantial reconstruction errors.

Several key discrepancies exist between DRRs and real X-rays, one of the most significant being scattering. Scattering is prominent in real X-rays but absent in DRRs, which simulate only primary rays without accounting for scattered photons. Photon scattering due to tissue interactions introduces noise, blurring, and low-frequency artifacts that reduce contrast, especially in soft tissue areas. These artifacts are most noticeable near dense anatomical structures or equipment, where multiple photon scatterings occur before reaching the detector. In head and neck imaging, these effects are somewhat reduced due to lower tissue density and volume, but in systems like ExacTrac, the lower-energy X-rays increase scattering, producing a fogging effect that can blur anatomical boundaries.

Additional discrepancies arise from calibration differences, noise, and resolution disparities. Real X-rays are affected by system noise, movement artifacts, and hardware imperfections, while DRRs, generated using trilinear interpolation with the Siddons algo-

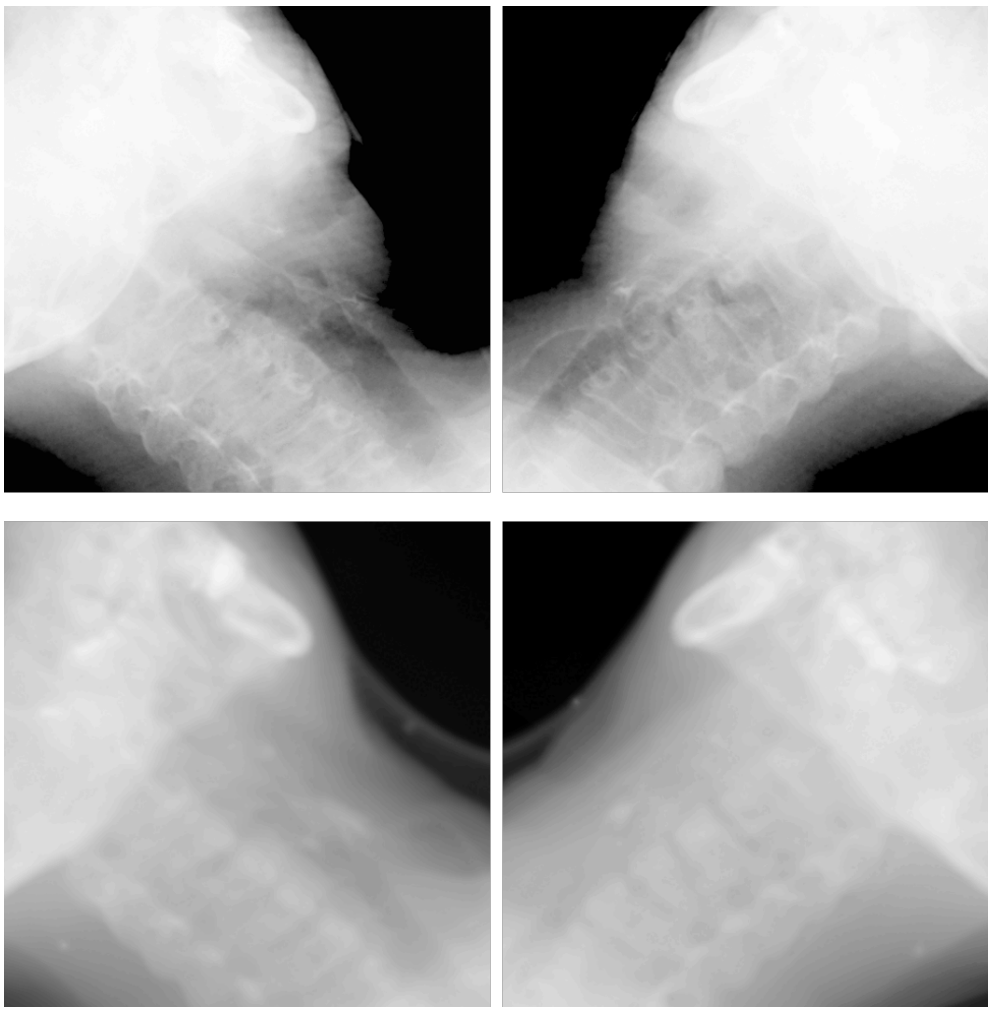

Figure 5.6: Comparison of DRRs generated from the virtual CT of the day (CT deformed to match CBCT, with real correction shifts applied)(bottom) with real X-rays(top).

rithm, have inherently lower resolution and sharpness due to voxel spacing (around 1–2 mm).

Moreover, DRRs capture a broader range of anatomical features, including soft tissues, whereas real X-rays focus more on bones, especially in ExacTrac, where settings are optimized for bony structures. These discrepancies are further amplified if the DRR generation energy spectrum does not perfectly match the the one of the real X-ray system.

Mapping Network

Accurate 3D reconstruction requires addressing discrepancies between DRRs and real X-rays, which arise due to scatter, energy differences, and resolution limitations. One way to bridge this gap is by simulating scatter during DRR generation, making DRRs more representative of actual X-ray conditions. Alternatively, scatter can be removed directly from real X-ray images. Refinements like enhanced energy models and super-resolution techniques further align DRRs with real X-rays, enhancing reconstruction accuracy.

Scatter correction is crucial for reducing image artifacts, especially in CBCT imaging. Traditional methods, such as Niu et al.'s multi-step algorithm [Niu, 2010], use deformable image registration to align planning CT with CBCT, creating a virtual CT from which DRRs are generated to correct scatter in CBCT projections. This approach also addresses low-frequency artifacts like beam hardening and kV differences, showing benefits in dose accuracy and OARs delineation [Kurz, 2016].

Zöllner et al. [Zöllner, 2017] proposed a similar approach, the Scatter-Correction Algorithm (SCA), which decomposes scatter and beam-hardening effects by subtracting DRRs from scaled CBCT projections, followed by smoothing and reconstruction. This method, closely matching Monte Carlo simulations in accuracy, offers effective scatter correction with lower computational requirements.

However, both approaches assume an initially accurate CBCT reconstruction to enable effective deformable registration of the planning CT, so to disentangle the scatter effect. This is not possible with 2 projections as there is too much ambiguity.

Deep learning offers a promising alternative for translating DRRs into more realistic X-rays. For instance, ScatterNet [Hansen, 2018] and DeepDRR [Unberath, 2018] demonstrated that CNNs can effectively learn scatter effects, outperforming traditional kernel methods while maintaining low computational demands. The network estimates Rayleigh scatter by being trained on pairs of DRRs and realistic X-rays generated through Monte Carlo simulations. Poisson noise is also added to simulate X-ray noise, closely resembling real-world imaging conditions.

To address multiple effects simultaneously, we can train a mapping model to directly learn scatter, noise patterns and artifacts from paired DRR-X-ray datasets. This approach would also allow backpropagation for end-to-end optimization of the reconstruction process.

Techniques like CycleGAN, used in X2CT-GAN [Ying, 2019], and dense residual networks, used in XraySyn [Peng, 2021], have shown promise for 2D-to-2D translation. How-

ever, these methods have primarily focused on qualitative results or 2D generation and have not been thoroughly tested for real 3D reconstruction using paired data.

With access to paired and aligned datasets, supervised learning approaches are expected to outperform unpaired methods like CycleGAN.

Training

To leverage this, we trained both a U-Net and an Attention U-Net [Oktay, 2018] to map between DRRs and X-rays. We tested both directions: from X-rays to DRRs and from DRRs to X-rays.

Our model takes two concatenated views as input and is trained using a combination of Mean Absolute Error (MAE) and perceptual loss functions, weighted at 1 and 0.1, respectively, to optimize both reconstruction accuracy and visual fidelity. The models were trained for half a day, with early stopping implemented at 120 epochs. The training used a learning rate of 3×10^{-4}, the Adam optimizer, and a batch size of 16.

Dataset

For this study, we used our longitudinal dataset of CT and CBCT scans from the CLB cohort, from which we extracted Exactrac images, configuration files for rendering matrices and corrections. We extracted clinical shifts from 212 patients across up to 35 treatment fractions per patient. These shifts, provided in the IEC coordinate system, were linked to the patients' orientation in the HFS position.

All longitudinal pairs of CT and CBCT scans were utilized, with an average of 14.5 CBCT scans per patient. Corrections were applied to generate virtual CTs aligned with each patient's current anatomy. From these virtual CTs, DRRs were created and paired with X-rays. As mentioned earlier, CBCT scans were typically acquired weekly when matched with ExacTrac registration or daily otherwise. We specifically selected cases where CBCT scans were available to create paired DRRs and X-rays.

This process resulted in 2,725 pairs of DRRs and X-rays from 188 patients. The intrapatient longitudinal changes in patient positioning and anatomy over time introduced diversity, serving as natural data augmentation. From the final dataset, 152 patients were used for training (with 2,173 pairs), 18 patients for validation (with 274 pairs), and 18 patients for testing (with 278 pairs).

Results

The quantitative evaluation of these models, measured by PSNR and SSIM, is shown in Table 5.1 for the Attention U-Net. Results show that mapping from X-rays to DRRs achieves better accuracy, likely due to the task's focus on removing noise and scatter rather than introducing additional effects. The attention-based U-Net outperformed the standard U-Net in both tasks. This model achieved a 40% improvement in PSNR compared to the initial discrepancy between X-rays and DRRs, demonstrating its effectiveness in reducing the gap between the two and improving translation quality.

Table 5.1: Quantitative evaluation of mapping with Attention UNet between DRRs and X-rays using PSNR and SSIM.

Mapping	**Method**	**PSNR**	**SSIM**
DRR $\rightarrow$ X-ray	MAE	19.6	0.87
	MAE + Perceptual	19.6	0.87
X-ray $\rightarrow$ DRR	MAE	23.0	0.88
	MAE + Perceptual	23.5	0.90

Figure 5.7 shows visual validation results from our best model, displaying the X-rays, the predicted DRRs alongside the targets DRRs for comparison.

The predicted images appear smoother, with better calibration and extended tissue visibility. However, some fine-scale details do not perfectly match the target DRRs, indicating room for improvement. The ultimate measure of the mapping network's quality will be its performance in direct 3D reconstruction from real biplanar X-rays. This will reveal whether the translation is accurate enough to enable reliable reconstruction and clinical application. We provide such trial in the final section 5.2.5.

5.2 Adapting Our Methods to Real Biplanar Systems

We aim to integrate our translation of X-rays into DRRs that we can geometrically generate to match real biplanar systems. This adaptation will allow us to apply our reconstruction methods to actual biplanar systems. However, this integration presents additional challenges that must be addressed to ensure seamless compatibility with real-world systems.

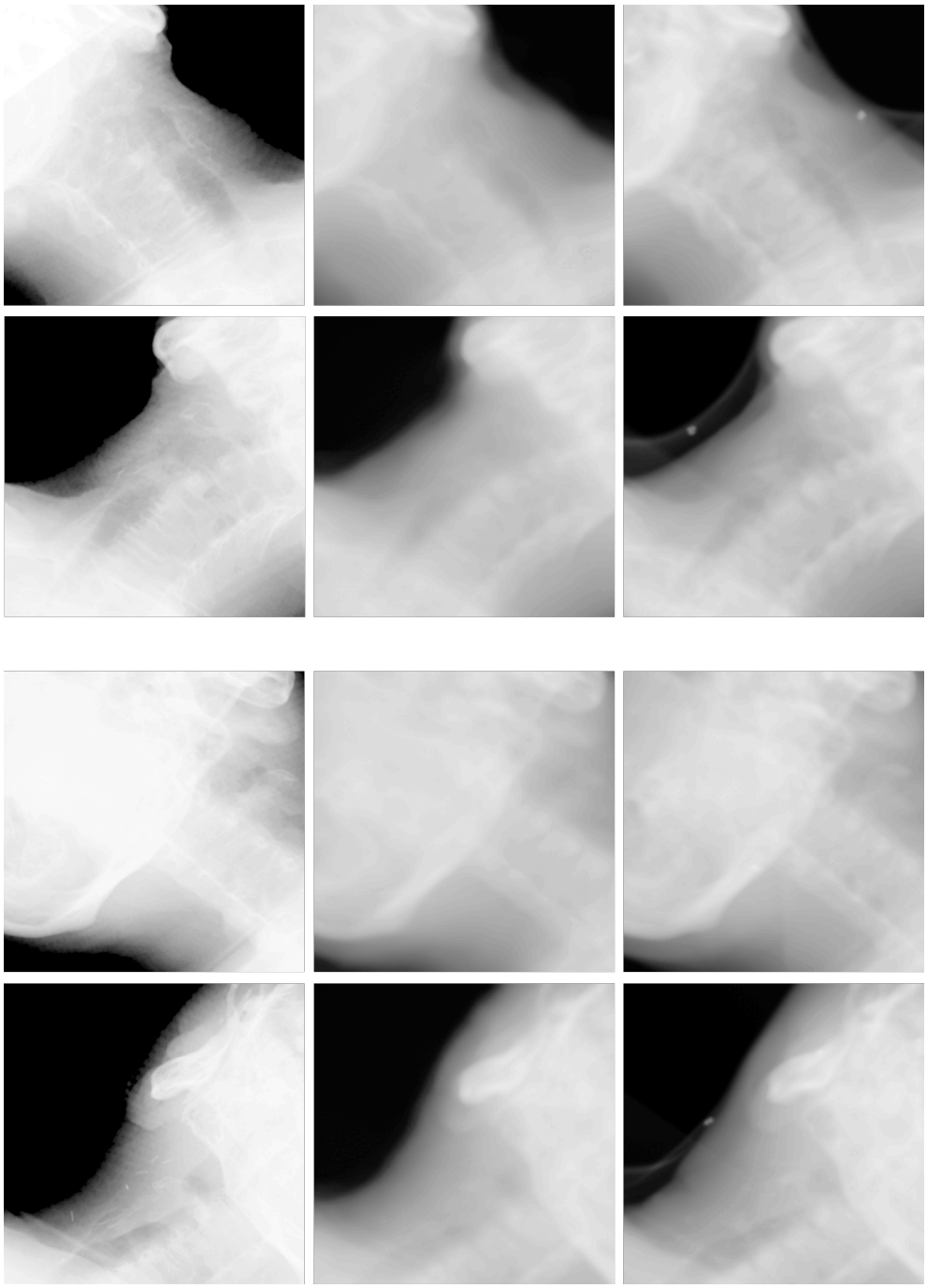

Figure 5.7: Prediction results from the mapping network transforming X-rays to DRRs for two validation cases (top and bottom). Each set includes the original X-ray (left), the model's prediction (middle), and the target DRR (right).

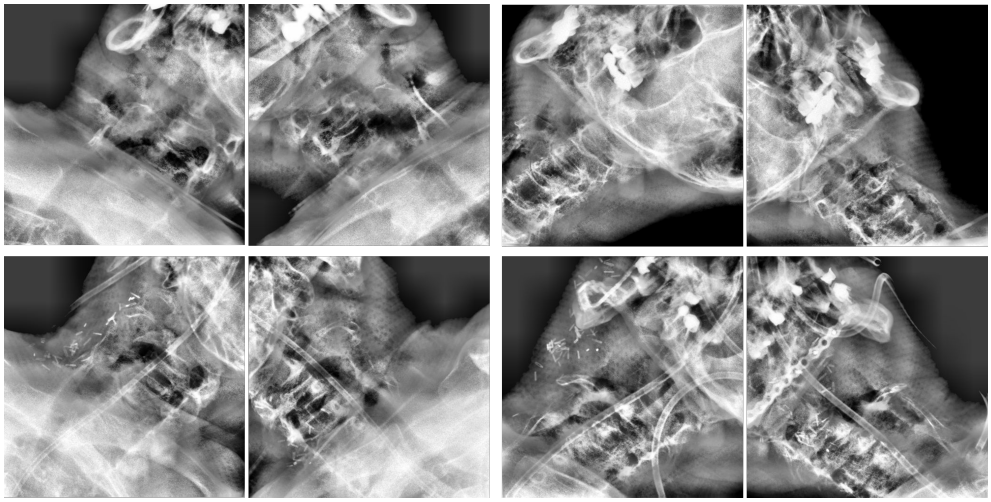

Figure 5.8: Examples of pairs of X-rays highlighting variations in isocenters and regions of interest in head and neck imaging, from the skull base to the upper lungs, with potential lateral shifts based on tumor location.

5.2.1 Challenges of Clinical Reality

Our models, X2Vision and XSynthMorph, were initially developed to demonstrate reconstruction potential specifically for the central head and neck region. However, in clinical practice, various factors come into play. Biplanar imaging focuses on the isocenter of the PTV, and tumors can extend throughout the entire head and neck region. This leads to a wide range of captured regions of interest.

Figure 5.8 illustrates examples of different isocenters and regions of interest in head and neck imaging. Real X-rays may capture areas from the top of the lungs, crossing the clavicles, up to the skull base, with lateral shifts depending on tumor location.

Also, oblique, non-coplanar X-rays intersect multiple cranio-caudal slices, capturing more tissue and introducing greater ambiguity than face-profile images. The narrow field of view provides only partial supervision, as certain regions are visible in just one projection, reducing the overlap between projection zones and leaving some 3D slices unsupervised. Additionally, dosimetry simulation for adaptive radiotherapy may require reconstructions that extend beyond these restricted fields of view.

Reconstruction from biplanar systems needs to consider these complexities, necessitating a larger reconstruction area and the ability to handle variability. To address this, we developed generative and deformation models that encompass the entire head and neck region.

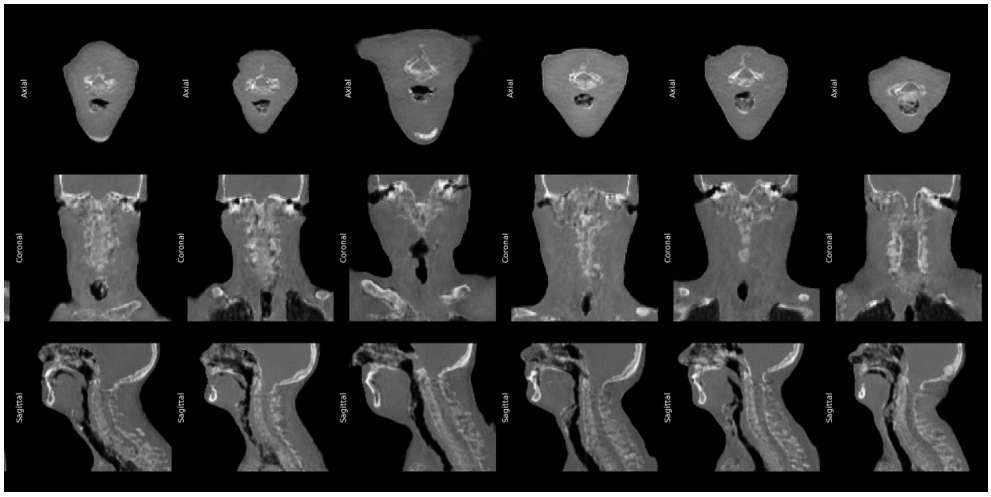

Figure 5.9: Several generations from the model on full head and neck region, axial(top), coronal(middle), and sagittal(bottom) slices are represented.

5.2.2 Generative Model

We trained a new generative model on a larger scale to cover the entire head and neck region, offering several key advantages for clinical applications. This model provides complete head and neck coverage, allowing for robust adaptation to varying FOVs and the ability to handle partial FOVs, making it well-suited for real clinical scenarios.

Dataset and Training

We used the same architecture as previously 4.4, that we trained over a period of three weeks. We also used the same dataset as previously 4.4, but this time focused on the full head and neck region, spanning from the top of the lungs and clavicles to the skull base, centered on the pre-segmented larynx. To ensure effective learning, we excluded cases with a limited FOV, resulting in a final set of 3,073 cases with complete head and neck coverage. The CTs were downsampled $112 \times 128 \times 112$ at a resolution of $2.1 \times 2.1 \times 2.1\,\text{mm}^3$ to fit GPU capacity.

Results

The generative model's performance was evaluated using FID we got 55 which is a bit more compared to the first model (46).

Figure 5.9 shows the diversity of realistic whole head and neck structures generated by the model.

However, the details in the generated images are less refined compared to the first model, with certain features like the cervical structures and tissues being less well reconstructed. Training this model was more challenging due to the larger and more complex anatomical variations. Considering we used the same complexity level in terms of network

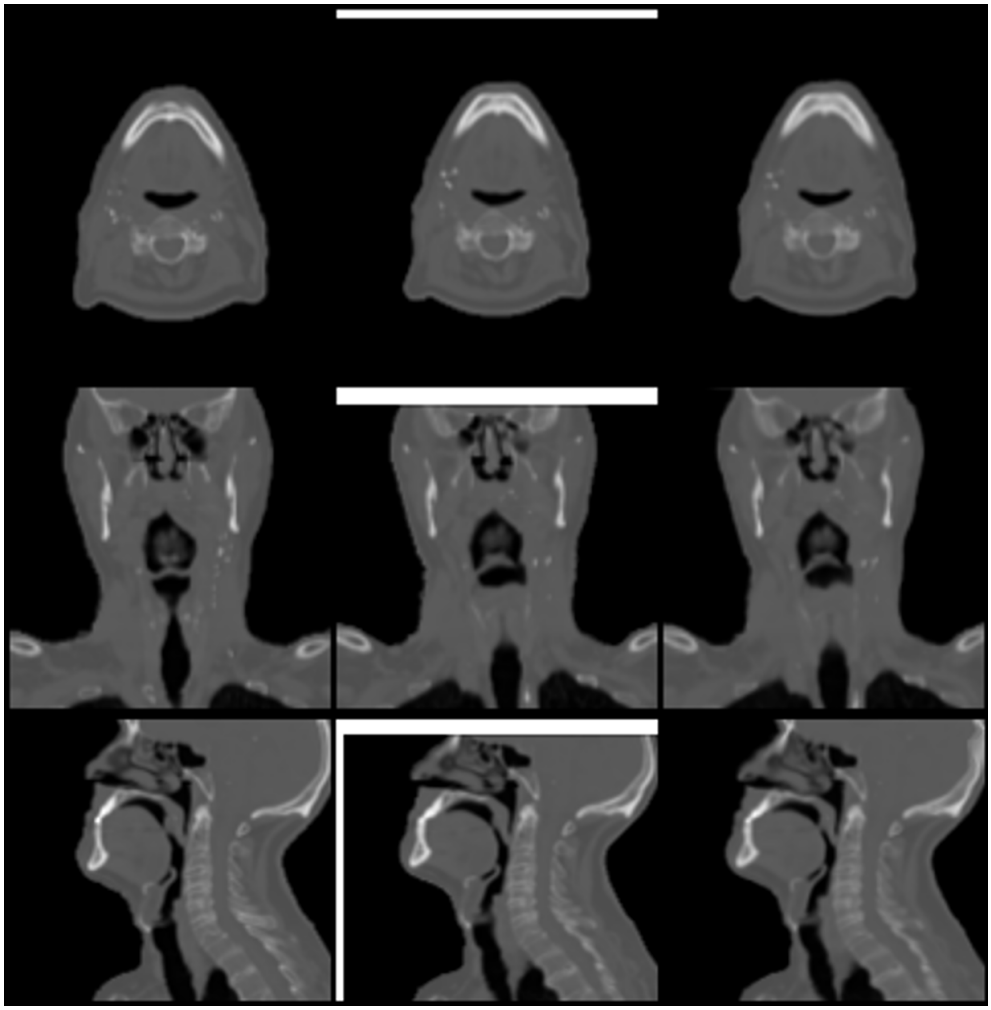

Figure 5.10: Validation example illustrating the planning CT (left), the deformed CT aligned with a later-stage CBCT with a slightly limited field of view as the target (middle), and the deformation result predicted by our model (right). This visualization highlights the effectiveness of our approach in accurately capturing anatomical changes over time.

architecture, more filters could potentially improve the learning of the manifold.

This model could adjust to varying poses and regions of interest by incorporating prior anatomical knowledge of the entire head and neck. However, for aligning with the varying poses, it is necessary to align the generation with the specific region of interest, which requires pre-registration as discussed in next section 5.2.4. We note that it is simpler for the GAN to learn a fixed manifold rather than dealing with positional changes during training.

5.2.3 Deformation Model

We also extended our deformation model to encompass the entire head and neck region. Using the same VoxelMorph framework as in XSynthMorph 4.4, we adapted it to handle full head and neck coverage.

Dataset and Training

The model was trained on planning CTs paired with corresponding CBCTs, even when the CBCTs provided only partial information. Compared to the initial XSynthMorph model 4.4, we expanded the dataset to include all possible longitudinal pairs, not just those associated with projections. This yielded 3,868 pairs from 371 patients, with an average of 11.1 CBCT scans per patient. The training set comprised 2,990 pairs from 314 patients, while the validation set included 507 pairs from 35 patients.

CTs were processed to match the resolution and size of the generated data ($112 \times 128 \times 112$ at $2.1 \times 2.1 \times 2.1\,\mathrm{mm}^3$). This approach eliminated the need for downsampling and upsampling during the reconstruction, which should reduce blurriness by removing interpolation steps.

The FOV in CBCT scans often targets specific areas, such as the upper or lower sections, based on the PTV, resulting in partial coverage of the head and neck region. To train effectively under these constraints, we focused on learning deformations within the available zones of the head and neck. A masking strategy applied the pixelwise loss only within the overlapping zones of CBCT and the full region. In addition, we used a gradient smoothing loss across the entire head and neck to ensure smooth deformation propagation, even in regions outside the CBCT's FOV. This approach allowed us to utilize all available longitudinal CT-CBCT pairs, capturing diverse anatomical poses and isocenters. Consequently, deformations learned from partial CBCTs generalized well across the head and neck region, ensuring realistic alignment throughout.

The model was trained with the same hyperparameters and strategy as previously defined but over an extended number of epochs of 1500, to account for the added diversity and complexity across the entire head and neck region.

Rigid transformations were not included, as varying poses will be handled through pre-positioning adjustments, as explained in 5.2.4.

Results

The validation MSE was 1×10^{-3}, which is comparable to the previous model. Extended training allowed the model to capture a greater range of deformations and complexities.

Figure 5.10 illustrates the model's deformable registration on a validation case, highlighting the alignment of the head and neck within a partial FOV and the slight extension of deformation.

5.2.4 Rigid Pre-Positioning

Since the generative and deformation models were trained on a fixed head and neck region centered around the larynx, accurate alignment with the region of interest is crucial. This alignment can be achieved through rigid 2D/3D pre-registration. We can either use ExacTrac's clinically derived registration corrections or optimize the six degrees of freedom (translation and rotation) in-house using gradient descent to minimize the NCC between the projections and X-rays.

Additionally, ExacTrac's pre-registration aligns the model based on the CT isocenter, which corresponds to the PTV. To ensure consistency, we need to align our models with this isocenter. We calculate the offset between the center of the head and neck and the PTV isocenter, applying a translation to align the generative model with the projection window.

After establishing registration—whether through ExacTrac or our custom method—we align the models with the visible region in the actual X-rays to maintain geometric consistency. For validation, we utilized Brainlab's registration system, which allows for direct comparison with CBCT and serves as a reliable reference for accurate alignment. While our in-house approach offers an end-to-end solution, it introduced slight discrepancies compared to CBCT alignment.

This process ensures that the models are correctly positioned within the anatomical region, facilitating accurate comparison with real X-rays for 3D reconstruction.

5.2.5 Reconstruction with Real Biplanar X-Rays

Our goal is to integrate all developed models to enable reconstruction using real biplanar X-rays. This approach combines the following components:

- **Generative Model:** This model maintains robustness across varying poses and partial fields of view, providing anatomical priors for the entire head and neck region, as outlined in Section 5.2.2.
- **Deformation Model:** This model adapts to anatomical changes throughout the head and neck region, as described in Section 5.2.3.
- **Rigid Pre-Positioning:** This process aligns all components in 3D for reconstruction, as detailed in Section 5.2.4.

- **Real Projector:** This component generates DRRs that match the FOV and geometry of real X-rays, as defined in Section 5.1.2.
- **Mapping Network:** This network translates X-rays into DRRs to reduce noise, enhance tissue visibility, correct calibration, and mitigate scatter effects, as explained in Section 5.1.3.

We now aim to minimize the following loss function, updated from Eq. 4.3 :

$$\boldsymbol{g}^* = \underset{\boldsymbol{g}}{\operatorname{argmin}} \sum_i \mathcal{L}_i \left(R_{\text{rigid}}(S(v^-, v(\boldsymbol{g}))), M(I_i) \right) + \mathcal{R}(\boldsymbol{g}) \tag{5.2}$$

where each term is defined as follows:

- $v(\boldsymbol{g})$: The generative model of volumes, parameterized by $\boldsymbol{g}$, which produces a volume $v(\boldsymbol{g})$.
- $S(v^-, v(\boldsymbol{g}))$: The spatial transformer that deforms the initial volume v^- using the generated volume $v(\boldsymbol{g})$.
- R_{rigid}: The rigid transformation that applies translation and rotation (6 DoFs) to align the reconstruction with the X-ray isocenter.
- $M(I_i)$: The mapping network that translates the real X-rays I_i into DRRs for comparison.
- $\mathcal{L}_i$: The loss term comparing the projections of the transformed volume with mapped DRRs $M(I_i)$, ensuring alignment between the reconstruction and the real X-rays.
- $\mathcal{R}(\boldsymbol{g})$: The regularization term applied to the generative model.

The adapted loss term $\mathcal{L}_i$, designed for real biplanar systems with partial fields of view, is defined as:

$$\mathcal{L}(v, I_i) = \lambda_2 \left\| A_{\text{Real}_i} \circ v - I_i \right\|_2 + \lambda_p \mathcal{L}_p(A_{\text{Real}_i} \circ v, I_i) \tag{5.3}$$

where A_{Real_i} is the operator projecting the volume v under view i, mimicking the oblique geometry and limited field of view of systems like ExacTrac (as defined in 5.1.2).

Optimization

The optimization was performed similarly to XSynthMorph, following the same warmup phase as defined in 4.4.5, and using the same hyperparameters outlined in 4.5.2.

However, the process took longer—around 2 minutes and 30 seconds. The DRRs were rendered to mimic real ExacTrac systems using upsampled volumes at a resolution of $1\,\text{mm}^3$ (up from $2.1\,\text{mm}^3$), which increased the computation time due to the larger volumes involved in ray tracing. Rendering directly from the learned resolution could significantly reduce this time, bringing it closer to the more acceptable previous inference time of around 1 minute.

Additionnally, the models were also trained on larger volumes of $112 \times 128 \times 112$, compared to the original size of $80 \times 96 \times 112$, leading to longer times for generation and deformation prediction. Future optimizations could include network distillation techniques or improvements in the optimization process for better efficiency, along with other strategies discussed in 3.6.

Translating X-rays to DRRs as we do, rather than the reverse, actually helps to keep the reconstruction process efficient. This translation only needs to be done once before optimization, rather than converting generated DRRs to X-rays at each iteration.

Also, the rigid registration R_{rigid} can be either fixed or free. Allowing it to vary during optimization may enhance alignment but complicate the process. For now, we use a fixed approach, as the generative and deformation models are designed to adapt to minor variations.

Results

Unfortunately, the reconstruction using real X-rays has not been successful enough to be presented at this time. Further investigation is needed to address the reasons behind this limitation, most likely related to the mapping between X-rays and DRRs that we will explain in next section 5.2.5.

We will focus on presenting results obtained with biplanar DRRs and discuss the challenges and limitations of the reconstruction process.

To evaluate our complete approach, we utilized the same longitudinal test set as in previous experiments. This test set consists of 18 patients from the CLB cohort, selected from the original group of 70. These patients were included in the mapping network test to ensure an unbiased evaluation.

Figure 5.11 shows an example of test reconstruction result using generated biplanar DRRs with partial ExacTrac geometry, compared against the target CT deformed on CBCT with applied corrections. The DRR-based reconstruction, despite relying on projections with a limited FOV, shows promising alignment with previous results in simulated-real clinical settings. However, we do observe some reduction in accuracy compared to earlier findings 4.5.4.

Several factors contribute to these limitations. Due to the small FOV, not all regions

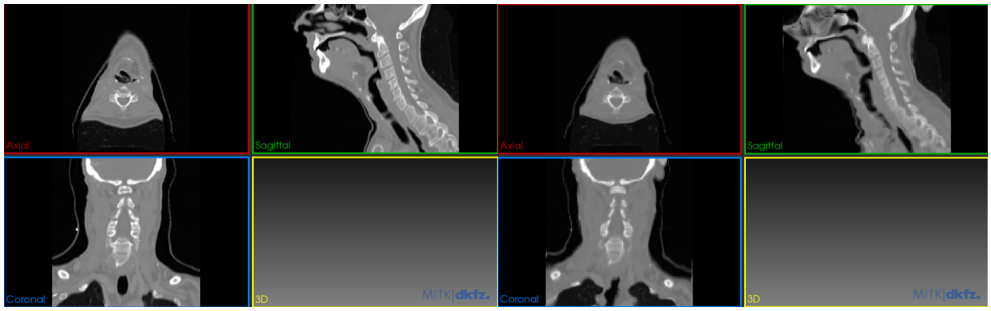

Figure 5.11: Comparison of reconstruction using realistic biplanar DRRs (right) and the target (CT deformed on CBCT with 6D registration corrections applied) (left).

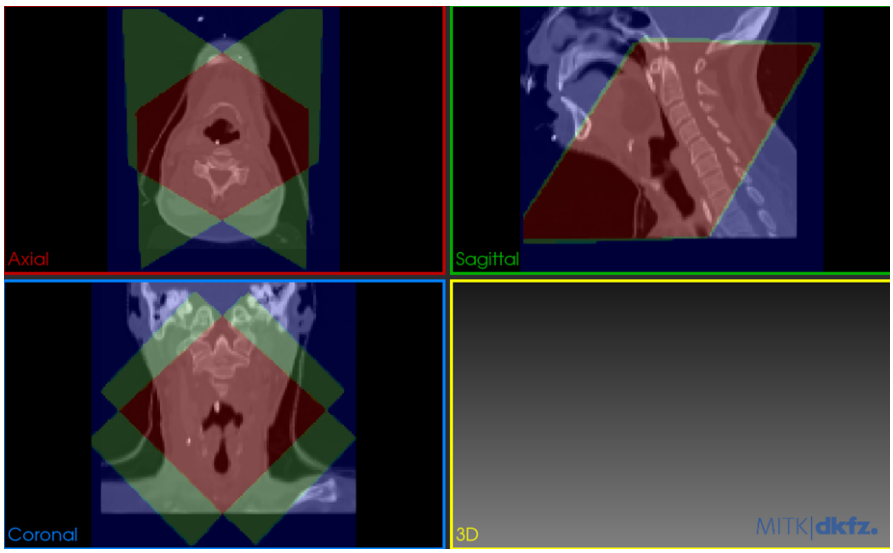

Figure 5.12: Visualization of partial FOV coverage. Red areas represent regions fully supervised by both X-ray views, while green areas indicate partial supervision by only one view, which complicates reconstruction in these less-supervised zones. The oblique projection angles (53° from the floor and 62° between X-rays) and limited FOV further increase ambiguity.

are intersected by X-rays, leading to partial supervision. The reconstruction focuses on minimizing the loss in the supervised zone, with the main reconstruction occurring in this area.

Figure 5.12 illustrates the areas covered by both X-rays and their intersection. Red regions indicate full supervision by both views, while green areas are only partially supervised by one projection, complicating reconstruction in these less-supervised zones. The projections, separated by 62° degrees instead of 90°, further reduce the potential for tissue disentanglement. Additionally, the oblique angle introduces increased ambiguity by intersecting more tissue layers.

By cropping to the actual FOV, as shown in Figure 5.13, we can focus on the areas crossed by X-rays, where main reconstruction is expected.

At this stage, we can present only qualitative results from our full approach, and

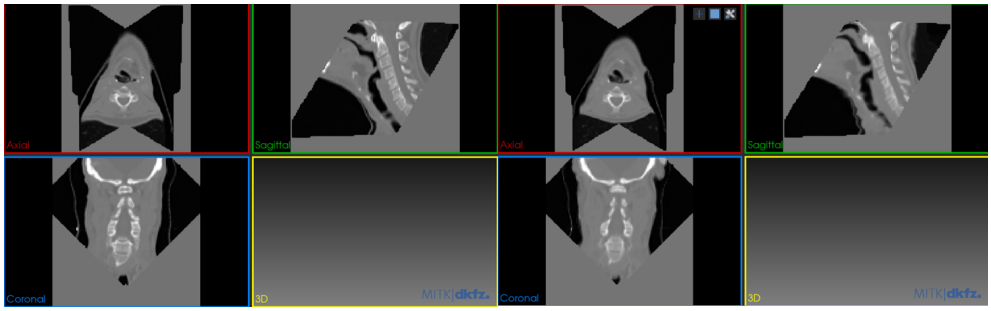

Figure 5.13: Comparison of the reconstructed image (right) and target image (left), each cropped to the partial FOV to focus on the regions covered by X-rays. Cropping to the actual FOV highlights areas with primary supervision and reconstruction.

additional quantitative analysis and experimentation are necessary to move toward clinical translation. Yet, the initial results are promising.

Additionally, further research is needed to evaluate the model's ability to reconstruct regions outside the X-ray field of view. Utilizing generative and deformation priors across the entire head and neck could facilitate realistic extensions beyond the current field. The generative model offers a realistic anatomical framework that could support accurate deformations and transitions throughout the head and neck, potentially allowing for a complete reconstruction both within and beyond the supervised area.

Challenges with Extending Reconstruction to Real X-Rays

The limited success in extending reconstruction to real X-rays likely stems from an imperfectly learned domain gap between DRRs and real X-rays. This discrepancy may result from structural misalignment between X-rays and DRRs, where small pixel shifts across slices can cause the mapping network to misinterpret tissue attenuation, leading to blurred or even missing structures.

Several factors contribute to these alignment challenges. Random patient movement between X-ray and CBCT acquisitions, particularly as anatomical changes like tissue loss occur during treatment, can cause misalignment. Restricting the training dataset to the initial three weeks of treatment—before significant anatomical shifts occur—could improve alignment accuracy. Additionally, minor inaccuracies in applied correction shifts may have led to imperfectly aligned training pairs, highlighting the need for thorough validation. Developing a pipeline that integrates rotational and translational corrections into the mapping network could enhance the alignment between X-rays and DRRs.

Also, differences between contrast-injected planning CTs and non-injected images acquired during treatment can alter projections. Artifacts from unmodeled elements, such as surgical tools or masks that appear in CBCT but are absent in DRRs, create further mismatches. Access to the real X-ray spectrum could help in accurately mimicking tissue attenuation.

The limited FOV of CBCT adds another challenge, as projections only partially cover

some regions, leading to edge effects when estimating deformations. A small FOV in training or testing can affect the comparison with real projections.

Deformable registration also has limitations, especially when faced with tissue loss or significant deformations that introduce artifacts and noise. A more accurate approach, such as a CBCT-to-CT mapping network, could improve alignment prior to reconstruction.

Scattering and noise in real X-rays present additional difficulties, as they introduce artifacts that are challenging to fully correct. Scattering patterns vary based on anatomy and FOVs, and while DeepDRR [Unberath, 2018] has shown that scattering and noise patterns can be learned with perfectly paired DRRs and Monte Carlo-simulated X-rays, applying this approach to real X-rays remains challenging.

Lastly, a network designed for direct 3D-to-2D translation, rather than focusing only on 2D-to-2D translation, could help establish stronger correlations between anatomical structures and the rendered X-rays. This approach could reduce ambiguities associated with 2D projections, which inherently lose a significant amount of structural information.

5.3 Conclusion and Discussion

This work presents a novel framework for integrating 3D CT reconstruction from biplanar X-rays within real clinical systems like ExacTrac. This marks the first exploration of 3D reconstruction for radiotherapy guidance directly within real biplanar systems, leveraging generative and deformable models trained on anatomical priors over the whole head and neck. Key components of the framework include calibrated projection generation aligned with ExacTrac, translations between X-rays and DRRs, and pre-registration via 2D/3D optimization to enable localized reconstructions adaptable to patient variability and clinical field-of-view constraints.

Current results using real X-rays are still preliminary, with limitations in accuracy, primarily due to differences in alignment, scattering and noise between real X-rays and DRRs. Further improvements in the translation model between DRRs and X-rays are needed to address these discrepancies. Achieving robust 3D reconstructions with real X-rays comparable to DRR-based reconstructions, presented here and in 4.5.4, would offer a transformative, low-cost, and low-dose alternative to CBCT for adaptive radiotherapy.

Clinical Translation Future clinical applications of this approach will require extensive validation using real X-rays and quantitative evaluations. Key considerations for clinical translation include validating the method in both rigid and deformable registration contexts, both within and beyond the primary field of view. Dosimetric studies are essential to confirm the method's accuracy in reflecting patient anatomy for treatment planning and to evaluate its utility in adaptive radiotherapy settings. Achieving results comparable to those shown in Section 4.5.4 would provide strong support for adaptive radiotherapy, as discussed in the previous chapter 5.

To ensure clinical viability, a quality assurance process mirroring current CBCT protocols could be established. For instance, a weekly CBCT comparison with reconstructed images could verify alignment accuracy; if the reconstruction proves reliable, the remainder of the week's imaging could rely on biplanar X-rays. This routine would allow continuous monitoring and ensure that anatomical changes align with the model's reconstruction capabilities.

Evaluating the robustness and uncertainty of this approach, especially in dose estimation and registration accuracy, is important. Ensuring that uncertainties remain within clinically acceptable limits for both dosimetry and anatomical registration would establish the method as sufficiently reliable for clinical use, with routine QA to sustain accuracy.

Out-of-Field Extension Accurate dosimetric calculations require extending the reconstruction beyond the current limited field of view, which presents additional challenges. The existing generative and deformation models, trained on larger anatomical region, show great potential in this context. Integrating rigid registration techniques, as described in 3.5.3, along with extended anatomical models and biomedical constraints, could enable smooth transitions into regions outside the primary FOV. This approach would provide a comprehensive foundation for dosimetric assessment across a broader area.

Adaptability of the Method This framework is adaptable to other anatomical regions beyond the head and neck and could extend to various biplanar systems. The generative and deformation models were designed to be unsupervised, allowing for a broad generalization across imaging systems. Calibration and DRR translation training for each specific system would suffice for method adaptation, offering a practical route for broader clinical adoption.

Additionally, high-quality pre-acquired 3D imaging data, such as planning CT or MRI, will remain essential, at least in the near future. Our method can utilize any pre-acquired volume with density information to generate projections. Pre-acquired MRI or CBCT data can be converted into CT-like images to provide density priors, thereby expanding the applicability of this approach.

Multi-Modal Integration for Enhanced Precision In the future, this framework could integrate additional modalities, such as surface-guided radiotherapy (SGRT), which captures 3D surface data for patient positioning using cameras or infrared imaging, as seen in ExacTrac Dynamic [AG, 2024] and VisionRT [Vision RT, 2024] systems. Combining external surface contours as an additional constraint for 3D reconstructions from X-rays would create a more robust and informed multi-modal solution for image-guided radiotherapy.

Real-Time Adaptive Radiotherapy A forward-looking direction involves real-time adaptive radiotherapy. By integrating this framework, widely fasten, into online systems, near real-time adjustments to the treatment plan could be made using continuously updated anatomical information. X-rays would be acquired every few seconds during treatment, as in systems like ExacTrac Dynamic or CyberKnife [Accuray Incorporated, 2024], with 3D reconstructions performed on-the-fly. The treatment plan would be updated dynamically, similar to practices in MR-Linac systems. This approach would be particularly useful for treatments involving respiratory motion or other dynamic anatomical changes.

Chapter 6

Conclusion

The pursuit of precision in radiation therapy is of paramount importance, especially in the treatment of head and neck cancers where complex anatomies and the proximity to vital structures demand exceptional accuracy. This thesis addresses a highly ill-posed yet crucial problem in this domain: reconstructing high-quality three-dimensional anatomical images from minimal, low-dose imaging data—in this case, biplanar X-rays. The work presented introduces innovative methods that not only overcome the limitations of current imaging techniques but also pave the way for transformative advancements in adaptive radiotherapy.

At the core of this research is the development of two novel unsupervised deep learning frameworks, X2Vision and XSynthMorph, designed to tackle the ill-posed problem of 3D reconstruction from limited two-dimensional projections. X2Vision leverages a generative prior to capture the anatomical manifold of the head and neck region. By constraining the solution space to anatomically plausible structures, it effectively reduces the ambiguity inherent in reconstructing 3D images from biplanar X-rays. This method demonstrates that, within the learned anatomical manifold, even as few as two projections can suffice to recover a close approximation of the true 3D anatomy.

Building upon this foundation, XSynthMorph introduces deformation priors by incorporating patient-specific pre-treatment CT. This integration allows the model to account for patient-specific anatomical variations and non-rigid deformations such as tumor regression and tissue changes over the course of treatment. The method jointly optimizes the anatomical generation and deformation, achieving unprecedented accuracy in reconstructing patient anatomy. These theoretical advancements highlight the power of combining learned anatomical priors with patient-specific data to solve highly ill-posed inverse problems in medical imaging. They demonstrate that the solution space can be significantly constrained through deep generative models and deformation fields, enabling accurate reconstructions from minimal input data.

From a clinical perspective, the methods developed in this thesis have significant im-

plications for adaptive radiotherapy. By enabling high-quality 3D reconstructions from low-dose biplanar X-rays, the need for frequent, high-dose imaging modalities like CBCT is mitigated. This reduction in cumulative radiation dose is particularly beneficial for patients requiring daily imaging. The accurate 3D reconstructions facilitate precise patient positioning and alignment, comparable to current CBCT-guided methods. This precision is crucial for targeting tumors effectively while sparing healthy tissue. Furthermore, the ability to account for anatomical changes over time allows for dynamic adaptation of treatment plans. Preliminary evaluations have shown promising results in structure retrieval and dosimetry analysis, which are essential for dose accumulation and potential replanning. Additionally, the methods offer a low-cost alternative to expensive imaging systems, making advanced radiotherapy techniques more accessible, especially in low-resource settings and emerging countries. Adaptations of the methods for integration with existing clinical biplanar X-ray systems have been explored, considering practical factors like limited fields of view and imaging noise, paving the way for seamless incorporation into current radiotherapy practices.

The advancements presented in this thesis open new horizons for radiation oncology. Envisioning a future where rapid 3D reconstructions enable real-time treatment adaptations, similar to MR Linac systems but with significantly lower costs and broader accessibility, is now within reach. This would allow for continuous adjustments to treatment plans in response to anatomical changes, improving outcomes and reducing side effects. Combining these reconstruction methods with other non-invasive imaging modalities, such as surface-guided radiotherapy or even integrating predictive models linked to genomics, could further enhance accuracy. This fusion of data would limit the degrees of freedom in reconstruction, ensuring minimal radiation exposure while maximizing precision. The ultimate goal is to deliver highly individualized treatments that adapt to each patient's unique anatomy and tumor characteristics over time, aligning with the broader movement towards precision medicine in oncology. Incorporating automation in segmentation, treatment planning, and dose prediction, driven by artificial intelligence, will streamline workflows and reduce the burden on clinical staff. Ensuring the reliability and robustness of AI-generated outputs will be critical, necessitating ongoing research into quality control and uncertainty quantification.

This thesis represents a foundational step toward revolutionizing adaptive radiotherapy through innovative imaging solutions. By successfully demonstrating the feasibility of reconstructing accurate 3D anatomical images from biplanar X-rays, we have challenged the conventional boundaries of medical imaging and opened the door to more patient-friendly, efficient, and accessible cancer treatments. While significant work remains to translate these methods fully into clinical practice—including extensive validation, addressing challenges of real-world data variability, and ensuring regulatory compliance—the potential benefits are immense. The methods hold promise not just for head and neck cancers but could be extended to other anatomical regions and cancer types, amplifying their impact. Ultimately, this research strives to enhance patient outcomes by improving the precision

of radiotherapy treatments while reducing associated risks and burdens. As we look to the future, we anticipate that continued advancements in deep learning, imaging technologies, and clinical integration will bring us closer to realizing the full potential of truly adaptive, personalized radiotherapy.

Chapter 7

Appendix : 3D Cerebral Vasculature Reconstruction from Biplanar DSAs

Two Projections Suffice for Cerebral Vascular Reconstruction [Cafaro, 2024a]

3D reconstruction of cerebral vasculature from 2D biplanar projections could significantly improve diagnosis and treatment planning. We introduce a novel approach to tackle this challenging task by initially backprojecting the two projections, a process that traditionally results in unsatisfactory outcomes due to inherent ambiguities. To overcome this, we employ a U-Net approach trained to resolve these ambiguities, leading to significant improvement in reconstruction quality. The process is further refined using a Maximum A Posteriori strategy with a prior that favors continuity, leading to enhanced 3D reconstructions. We evaluated our approach using a comprehensive dataset comprising segmentations from approximately 700 MR angiography scans, from which we generated paired realistic biplanar DRRs. Upon testing with held-out data, our method achieved an 80% Dice similarity w.r.t the ground truth, superior to existing methods.

7.1 Introduction

Digital Subtraction Angiography (DSA) plays an important role in the planning and treatment of neurovascular diseases providing surgeons with rich information about the brain angioarchitecture and hemodynamics [Ruedinger, 2021]. Although 3D MRA, CTA, or rotational DSA exist, 2D DSA remains the gold standard, due to its high resolution and clinical availability. DSA is commonly acquired as a set of biplanar anterior-posterior (AP) and lateral (L) projections of the vascular network [Settecase, 2021; Haouchine, 2021]. Unlike 3D rotational scanners, which are not suitable for real-time interventions, and sin-

gle DSAs, which are limited to simpler tasks, biplanar DSAs offer an optimal balance of speed, anatomical constraints, cost efficiency, and reduced radiation exposure.

Yet, projection onto 2D images causes vessel overlap, which makes it difficult for surgeons to confidently localize lesions, understand their shapes and morphologies, or distinguish between vessels feeding and draining malformations when the number of 2D views is limited [Settecase, 2021; Haouchine, 2021]. Thus, 3D reconstruction becomes critical.

Reconstructing cerebral vasculature from biplanar projections is a heavily ill-posed problem. The dense and intricate arrangements of blood vessels overlap and intertwine onto 2D projections, raising major ambiguities. While few attempts have been made to tackle this challenging problem, most of them focus on simpler vascular structures, like main coronary arteries, and typically require manual adjustments for vessel endpoints and bifurcations. To obtain more complex reconstructions, other techniques rely on pre-existing 3D models of patient's vasculature to add constraints or simulate flow [Copeland, 2010]. However, the availability of 3D imaging cannot be guaranteed in clinical practice. A non-learning approach [Frisken, 2022] relies on structural and temporal constraints but requires perfect tedious semi-manual annotations of segmented vessel centerlines from DSAs, limiting its use for real-time intervention. Alternatively, various deep learning techniques have been proposed, including self-supervised approaches [Zhao, 2022], Neural Radiance Fields (NeRF) techniques [Maas, 2023], denoising approach [Wu, 2023] and Generative Adversarial Networks (GANs) [Zuo, 2021]. These models typically aim at learning a prior to disambiguate DSAs and performing direct prediction. However, none of these techniques reached a good level of performance when only 2 projections are available. 3D backprojected volumes offers geometrical cues for reconstruction, albeit as noisy and ambiguous representations of the actual volume. Unlike the Denoiser approach [Wu, 2023], we show that a deep learning network can significantly clarify these volumes by learning priors on vascular patterns, resulting in closely matching reconstructions.

Contribution.

We propose in this paper a novel method for 3D reconstruction of DSA from only two projections. Our method follows a two step process; the first step involves a disambiguating reconstructor from back-projected volumes, built upon the Denoiser model. We enhanced the model with improved architecture and design to tackle the inherent complexities of the task. In the next step, we refine our initial predictions to find the Maximum A Posteriori (MAP) estimate given the projections. This refinement occurs through iterative optimization on a voxel grid, starting from our preliminary estimation of the vasculature. To improve the vascular network's structural integrity and connectivity, we introduce a connectivity prior inspired by Ising prior [Cipra, 1987]. We conduct several experiments to benchmark our method against the existing state-of-the-art and show that our method delivers a high level of reconstruction accuracy, closely matching the target vasculatures. It marks a significant improvement over existing techniques, suggesting that as few as two

projections might be sufficient for disambiguating structures for accurate 3D reconstruction. Our method is a promising approach, paving the way for validation on real data and potential clinical translation.

7.2 Method

Our proposed methodology uses a two-step pipeline, as illustrated in Figure 7.1. First, a 3D U-Net predicts an initial 3D vasculature from the back-projected volume. Second, Maximum A Posteriori estimation is employed to refine the initial 3D model with a newly introduced connectivity loss that encourages closing.

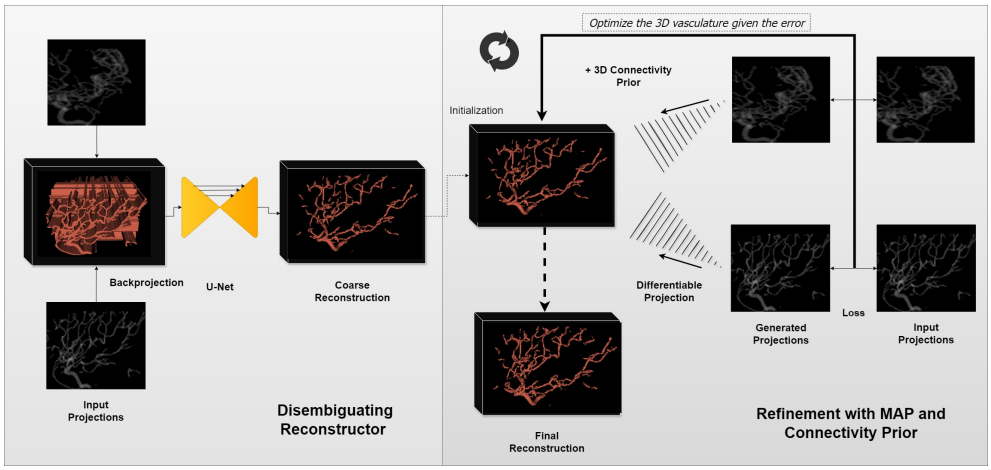

Figure 7.1: **Our pipeline.** Initially, we employ a 3D U-Net to generate a preliminary vasculature model from back-projected volumes. This model is subsequently refined through Maximum A Posteriori estimation, with the introduction of a connectivity prior, aimed at enhancing the model's structural cohesiveness and closure.

7.2.1 Disambiguating Reconstructor

Our goal is to address the challenging and ill-posed problem of reconstructing 3D brain vasculature from biplanar projections (I_0, I_{90}), acquired simultaneously on a bi-plane scanner, available in most interventional radiology suites. To create an initial 3D model of the brain vasculature, we propose to employ a 3D nnUNet [Isensee, 2021] model denoted as U_θ parametrized by the weights θ. Specifically, we first convert 2D images into a 3D volume V_b using back-projection. Then, the 3D nnUNet aims to disambiguate the back-projection to create a filtered volume corresponding to a 3D brain vasculature.

Given the absence of real paired projection-3D volume datasets, we propose to train our 3D nnUNet model using synthetic projections from ground truth volumes, which were then converted into back-projected 3D volumes. Specifically, we model blood vessels within a 3D space, assigning them a binary map with an intensity value of 1. Then we simulate biplanar projection using Mean Intensity Projection along rays, corresponding

to summing up log attenuation coefficients. This approach mirrors the accumulation of attenuation that occurs as X-rays pass through vessels. Finally, we create a backprojected volume V^b by extending these 2D projections into 3D along their original rays. Note that projection and backprojection are fast and differentiable. Overall, this enables us to create a paired dataset of back-projected 3D volumes and ground truth GT predictions.

To optimize parameters θ of our model, we use a combination of Dice and cross-entropy loss:

$$\mathcal{L}(U_\theta(V_b), GT) = \lambda_D \cdot \mathcal{L}_{Dice}(U_\theta(V_b), GT) + \lambda_{CE} \cdot \mathcal{L}_{CE}(U_\theta(V_b), GT), \tag{7.1}$$

where λ_D and λ_{CE} are weighting coefficients, $U_\theta(V_b)$ represents the predicted reconstruction for a given back-projected volume V_b and GT denotes the ground truth 3D vasculature.

7.2.2 Refinement of 3D Vasculature with MAP Estimate.

Building on our initial prediction, we then proceed to a refinement phase. This phase focuses on fine-tuning a 3D voxel-grid initialized with our deep learning output, $V_{coarse} = U_\theta(V_b)$, to improve 3D reconstruction alignment with the original projections, as well as connectivity.

MAP Estimation.

The process employs MAP estimation, using the pseudo-probability V_{coarse} as our initialization, i.e. $V_0 = V_{coarse}$. Our goal is to iteratively adjust this volume toward a MAP estimate V^*, achieving an optimal vasculature configuration that matches observed projections while enforcing structural integrity and connectivity, leading to the following optimization problem:

$$V^* = \underset{V}{\operatorname{argmax}} \log \mathcal{P}(V|I_0, I_{90}) = \underset{V}{\operatorname{argmax}} \mathcal{L}(V, I_0, I_{90}) + \mathcal{R}(V) \tag{7.2}$$

Here, $\mathcal{P}(V|I_0, I_{90})$ denotes the posterior probability of the vasculature V given the observed projections I_0, I_{90}, where $\mathcal{L}(V, I_0, I_{90})$ and $\mathcal{R}(V)$ respectively represents the log-likelihood of the current estimate to the observed projections, and a regularization term that integrates a specially designed connectivity prior.

Connectivity Prior.

We introduce a connectivity prior inspired by the Ising model [Cipra, 1987] to enhance voxel interconnectivity and create a more cohesive vascular network. In the context of vasculature, continuity is expected within vessels, except at termination points. The proposed reguarlization term encourages neighboring elements to exhibit similar values for spatial coherence, while allowing for natural discontinuities at boundaries or edges. To achieve this, we formulate a loss function that acts as an energy function:

$$\mathcal{R}_c(V) = -\frac{1}{N}\sum_{w=1}^{W}\sum_{h=1}^{H}\sum_{d=1}^{D}\sum_{x\in\mathcal{N}(w,h,d)} V_{w,h,d}\cdot V_x,$$

where $N = W \times H \times D$ and $\mathcal{N}(w,h,d)$ represents the set of all 26 neighboring voxels of (w,h,d).

This formula calculates the negative sum of the product of each voxel with its 26 neighbors within the 3D grid, normalized by the number of points $N = W \times H \times D$. Minimizing this energy function encourages connecting voxels by growing connections and filling in gaps, thus improving structural integrity and coherence.

Optimization.

The refinement begins with V_{coarse} and the optimization aims to balance data fidelity with the regularization informed by our connectivity prior. Eq. 7.2 refines as :

$$V^* = \underset{v}{\operatorname{argmin}} \left\|A_0 \circ V - I_0\right\|_2 + \left\|A_{90} \circ V - I_{90}\right\|_2 + \lambda_c \mathcal{R}_c(V) \tag{7.3}$$

where the λ_c is fixed.

A_i are the projector operators under view i.

7.3 Experiments and Results

We evaluate our method for our target application, 3D vasculature reconstruction from biplanar projections. In this section, we introduce our dataset, our model architecture, and present both quantitative and qualitative comparisons with state-of-the-art methods. We also include an ablation study to evaluate the contribution of our regularization prior.

7.3.1 Dataset and Preprocessing

Voxel-based Binary Vasculature Maps Creation.

Given the scarcity of 3D vasculature segmentations, we developed a comprehensive approach to generate a substantial dataset for training our deep learning model. Utilizing the publicly available TubeTK [Aylward, 2002] dataset, we processed MRAs from 100 healthy patients, including 43 with detailed ground truth segmentations. To augment the dataset further, we trained a nnUNet model on these binary maps, achieving a validation Dice score of 0.75. Using this model we segmented additional MRAs from the publicly available IXI dataset [Hammersmith Hospital London,], which comprises 580 Time-of-Flight (TOF) MRAs of healthy patients.

Clinical Realism.

MRA typically images vessels in both brain hemispheres. In contrast, DSA is typically used to image one arterial branch at a time. Thus, DSA has 1) less vascular complexity than MRA and 2) vasculature restricted to one hemisphere. To mimic this clinical reality, we partitioned MRA images along the brain mid-plane. As a positive side effect, as we could use both hemispheres independently, this doubled the size of our dataset from 680 to 1360.

Computational Constraints.

Given the substantial size of vasculature volumes, we optimized GPU efficiency by downsampling all volumes to a resolution of $0.8 \times 0.94 \times 0.94$. We employed Signed Distance Fields (SDFs) to maintain the integrity of thin vessels during downsampling, preserving thin vessel structures and reducing artifacts, unlike binary images which can cause these structures to break apart or disappear. We used FastGeodis [Asad, 2022] with truncation at 14 and level at 0.03. Additionally, by identifying the minimal bounding grid for the vasculature, we standardized the volumes to $112 \times 80 \times 128$.

Model Training.

We created Digitally Reconstructed Radiographs (DRRs) from these volumes to serve as projections. These projections were then used to generate backprojected volumes as input for our model. These were resampled to a size of $128 \times 96 \times 128$ and Z-Score normalized. Our dataset was divided into $1,029$ training cases and 257 validation cases, split between different patients.

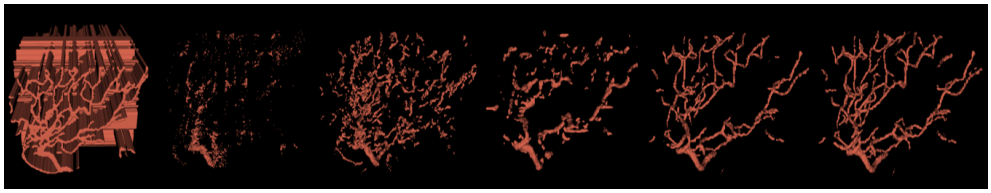

Figure 7.2: **Visual comparison of 3D reconstruction from biplanar projections by our model and baselines.** Backprojection produces highly noisy and ambiguous reconstructions. Grid optimization captures only broad structures and non-ambiguous segments. X2Vision offers slightly clearer reconstructions, capturing the main artery and some structures but remains fuzzy. The Denoiser reconstructs parts of the vessels but struggles with connectivity and complex branching areas. In contrast, our reconstruction closely mirrors the actual data, showcasing better well-defined patterns and accurately connected, complex vessel networks.

3D Reconstruction.

In the test phase, we utilized 70 distinct cases from the IXI dataset [Hammersmith Hospital London,] to assess our reconstruction methodology.

7.3.2 Implementation Details

Disambiguating Reconstructor.

Our model, built with PyTorch [Paszke, 2019] and tailored for the NVIDIA GeForce RTX 3090 GPU, adopts a nnUNet design with 6 encoding and 5 decoding blocks. It features asymmetric downsampling—5 times in larger dimensions and 4 times in the smallest—enhancing feature extraction across scales with channels increasing in the encoder (32, 64, 128, 256, 320, 320) and decreasing in the decoder (320, 256, 128, 64, 32). Skip connections improve information flow over simple encoder-decoders. Pooling mainly uses 2x2x2 kernel sizes, and convolutions are performed with 3x3x3 kernels. Running on a batch size of 3, the model starts with a 0.01 learning rate, using SGD with Nesterov momentum, a weight decay of 3e-5, and a polynomial rate scheduler. The loss function equally mixes dice and cross-entropy. Robustness is improved by extensive data augmentation, including spatial, noise, and contrast modifications. The model was trained in under a day for 500 epochs.

Optimization.

Our reconstruction optimizes a full-resolution voxel grid, initialized with the model's preliminary predictions from backprojected volumes. It is done on the same GPU using the Adam [Kingma, 2014] optimizer at a learning rate of $1\mathrm{e}{-1}$. Optimal weights ($\lambda_2 = 1$, $\lambda_c = 0.0002$) were determined through grid search. The process, taking 500 iterations, completes in about 20 seconds, but excluding connectivity loss cuts it down to 3 seconds.

Table 7.1: **Comparison with State-of-the-art.** Standard deviations in parentheses.

Method	Dice ↑	clDice ↑	Balanced HD ↓
Voxel-Grid	$0.22(\pm 0.11)$	$0.16(\pm 0.11)$	$1.81(\pm 0.26)$
X2Vision [Cafaro, 2023d]	$0.26(\pm 0.03)$	$0.20(\pm 0.03)$	$3.42(\pm 0.46)$
Denoiser [Wu, 2023]	$0.34(\pm 0.05)$	$0.28(\pm 0.06)$	$1.94(\pm 0.24)$
Ours (coarse)	$\mathbf{0.77}(\pm \mathbf{0.04})$	$\mathbf{0.75}(\pm \mathbf{0.04})$	$\mathbf{0.42}(\pm \mathbf{0.09})$
Ours (coarse w/ refinement)	$\mathbf{0.80}(\pm \mathbf{0.04})$	$\mathbf{0.78}(\pm \mathbf{0.04})$	$\mathbf{0.34}(\pm \mathbf{0.09})$

7.3.3 Results and Discussion

Baselines.

There are very few papers working on this specific task. We compared our approach against simple backprojection, learning-based supervised state-of-the-art Denoiser [Wu, 2023], voxel-grid optimization without prior, and unsupervised gan-based reconstruction model X2Vision [Cafaro, 2023d]. As no implementation of Denoiser was available, we reimplemented it.

Metrics.

For evaluation, we employed metrics including dice, cldice [Paetzold, 2019], and balanced Hausdorff distance (HD) [Aydin, 2021]. Compared to dice, cldice provides a balanced view, especially valuing the retrieval of branching patterns and minimizing the bias towards larger vessels.

3D Reconstruction from 2 Projections.

Figure 7.2 visually presents our reconstructions with baseline methods, while Table 7.1 details our quantitative results. Our approach significantly outperforms others, closely matching the actual vessel geometry and enhancing the precision of vessel proximity to targets, achieving impressive results.

Backprojection retrieves all possible locations of vessel presence, resulting in very noisy reconstructions with numerous false positives. Direct voxel-grid optimization, similar to our refinement process but without proper initialization, only reconstructs unambiguous areas and misses detailed structural nuances, leading to many false negatives. This underscores the complexity of reconstruction without prior knowledge.

X2Vision, which uses unsupervised GANs, struggles to learn the intricate and sparse nature of vascular structures. This weak prior allows it to identify the main artery but fails to produce continuous, realistic reconstructions. The Denoiser model, constrained by its simple architecture and lack of skip connections, only provides coarse vessel outlines and cannot capture complex vessel branching, showing limited improvement. We improved the Denoiser model by adopting a more complex architecture, including skip connections,

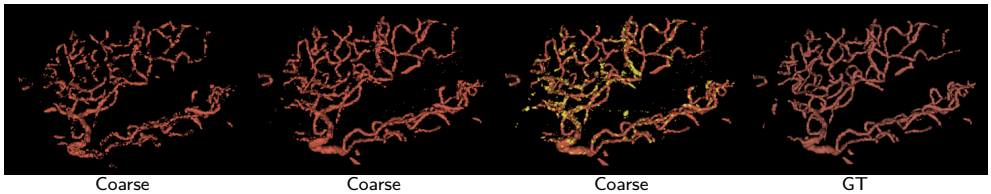

Figure 7.3: **Visual Ablation Study.** The initial results from our model are promising but exhibit gaps. Refinement with MAP enhances closure and refinement. Adding the connectivity prior further strengthens these improvements, allowing the model to closely replicate the intricate complexity of actual vasculature. The yellow shows the additional closing when introducing the connectivity loss.

Table 7.2: **Ablation Study.** Standard deviations are provided in parentheses.

Method	Dice ↑	clDice ↑	Balanced HD ↓
Grid	$0.22(\pm 0.11)$	$0.16(\pm 0.11)$	$1.81(\pm 0.26)$
Grid w/ Connectivity	$0.29(\pm 0.05)$	$0.23(\pm 0.05)$	$2.47(\pm 0.26)$
Coarse	$0.77(\pm 0.04)$	$0.75(\pm 0.04)$	$0.42(\pm 0.09)$
Coarse w/ Grid	$0.79(\pm 0.04)$	$0.76(\pm 0.04)$	$0.35(\pm 0.08)$
Coarse w/ Grid + Connectivity	$\mathbf{0.80}(\pm \mathbf{0.04})$	$\mathbf{0.78}(\pm \mathbf{0.04})$	$\mathbf{0.34}(\pm \mathbf{0.09})$

using a U-Net instead of a simple encoder-decoder, and combining dice and cross-entropy losses. We also introduced data augmentation and a refinement step. These enhancements significantly improved performance, leading to reconstructions that closely match the target vessel structures and branchings. Our refinement step further enhances the reconstruction. By introducing a MAP estimate refinement paired with a connectivity loss, we not only improve vessel connectivity but also refine vessel shapes and fill in missing structures, achieving more precise and closed vessel representations in the final volume.

Ablation Study.

Our ablation study, summarized in Figure 7.3 and Table 7.2, reveals that initial reconstructions may contain gaps, but MAP refinement and incorporating connectivity loss significantly enhance the quality. MAP refinement aligns structures with projections, while connectivity loss improves cldice scores by improving capture of vessel centerlines, outperforming grid optimization. It clarifies complex junctions and promotes interconnected high-probability voxels, filling gaps and enhancing structural integrity and coherence.

7.4 Conclusion and Future Work

We introduced a new method for the highly ill-posed 3D cerebral vascular reconstruction from biplanar DSAs. Our two-step approach starts with a disambiguating reconstructor, followed by refinement through MAP estimation and a connectivity prior. This method marks a notable advancement over existing methods, suggesting for the first time that two projections could effectively disambiguate complex vascular structures. Further improvements are expected with the integration of additional vascular properties. Due to GPU hardware capabilities, our work was limited to using downsampled volumes, restricting our method to vessels with diameters larger than 1-2mm. Additionally, we used synthetic DSA generated from automatically segmented MRA images due to the lack of paired MRA/DSA datasets. We are currently assembling such a dataset to validate our method on real data towards clinical translation.

Bibliography

[Abdal, 2019] Rameen Abdal, Yipeng Qin, and Peter Wonka. "Image2stylegan: How to embed images into the stylegan latent space?" In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 4432–4441.

[Accuray Incorporated, 2024] Accuray Incorporated. *CyberKnife Robotic Radiosurgery System*. https://www.accuray.com/cyberknife/. Accessed: 2024. 2024 (cit. on pp. 17, 160).

[AG, 2024] Brainlab AG. *ExacTrac Patient Positioning and Monitoring*. Accessed: [insert date here]. 2024 (cit. on pp. 17, 18, 134, 159).

[Aggarwal, 2018] Hemant K Aggarwal, Merry P Mani, and Mathews Jacob. "MoDL: Model-based deep learning architecture for inverse problems". In: *IEEE transactions on medical imaging* 38.2 (2018), pp. 394–405 (cit. on p. 55).

[Ahnesjö, 1989] Anders Ahnesjö. "Collapsed cone convolution of radiant energy for photon dose calculation in heterogeneous media". In: *Medical physics* 16.4 (1989), pp. 577–592 (cit. on p. 86).

[Arsigny, 2006] Vincent Arsigny, Olivier Commowick, Xavier Pennec, and Nicholas Ayache. "A log-euclidean framework for statistics on diffeomorphisms". In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2006: 9th International Conference, Copenhagen, Denmark, October 1-6, 2006. Proceedings, Part I 9*. Springer. 2006, pp. 924–931 (cit. on p. 116).

[Asad, 2022] Muhammad Asad, Reuben Dorent, and Tom Vercauteren. "FastGeodis: Fast Generalised Geodesic Distance Transform". In: *Journal of Open Source Software* 7.79 (Nov. 2022), p. 4532 (cit. on p. 170).

[Aydin, 2021] Orhun Utku Aydin, Abdel Aziz Taha, Adam Hilbert, Ahmed A Khalil, Ivana Galinovic, Jochen B Fiebach, Dietmar Frey, and Vince Istvan Madai. "On the usage of average Hausdorff distance for segmentation performance assessment: hidden error when used for ranking". In: *European radiology experimental* 5 (2021), pp. 1–7 (cit. on p. 172).

[Aylward, 2002] Stephen R Aylward and Elizabeth Bullitt. "Initialization, noise, singularities, and scale in height ridge traversal for tubular object centerline extraction". In: *IEEE transactions on medical imaging* 21.2 (2002), pp. 61–75 (cit. on p. 170).

[Bahdanau, 2014] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. "Neural Machine Translation by Jointly Learning to Align and Translate". In: *arXiv*. 2014.

[Bahmani, 2017] Sahand Bahmani and Justin Romberg. "Phase retrieval meets statistical learning theory: A flexible convex relaxation". In: *Foundations of Computational Mathematics* 17.6 (2017), pp. 1373–1398 (cit. on p. 60).

[Baker, 1985] R. Baker et al. "Simultaneous Iterative Reconstruction Technique (SIRT)". In: *Journal of Computer Assisted Tomography* 9.2 (1985), pp. 246–252 (cit. on p. 47).

[Balakrishnan, 2019] Guha Balakrishnan, Amy Zhao, Mert R. Sabuncu, John Guttag, and Adrian V. Dalca. "VoxelMorph: A Learning Framework for Deformable Medical Image Registration". In: 38.8 (2019) (cit. on pp. 115, 116, 118).

[Barker, 2004] J.L. Barker, A.S. Garden, K.K. Ang, et al. "Quantification of volumetric and geometric changes occurring during fractionated radiotherapy for head-and-neck cancer using an integrated CT/linear accelerator system". In: *International Journal of Radiation Oncology* Biology* Physics* 59.4 (2004), pp. 960–970 (cit. on p. 20).

[Bau, 2018] David Bau, Jun-Yan Zhu, Hendrik Strobelt, Bolei Zhou, Joshua B Tenenbaum, William T Freeman, and Antonio Torralba. "Gan dissection: Visualizing and understanding generative adversarial networks". In: *arXiv preprint arXiv:1811.10597* (2018) (cit. on p. 70).

[Beer, 1852] Beer. "Bestimmung der Absorption des rothen Lichts in farbigen Flüssigkeiten". In: *Annalen der Physik* 162.5 (1852), pp. 78–88 (cit. on p. 37).

[Behin Negareh Co, 2023] Behin Negareh Co. *X-ray Production and X-ray Tube*. https://behinnegareh.com/en/news/X-ray-Production-X-ray-Tube. Accessed: 2024-11. 2023 (cit. on p. 39).

[Beichel, 2015] R. R. Beichel, E. J. Ulrich, C. Bauer, A. Wahle, B. Brown, et al. "Data from QIN-HEADNECK". In: (2015) (cit. on p. 82).

[Beister, 2012] Marcel Beister, Daniel Kolditz, and Willi A Kalender. "Iterative reconstruction methods in X-ray CT". In: *Physica medica* 28.2 (2012), pp. 94–108 (cit. on p. 49).

[Bhadra, 2022] Sayantan Bhadra, Umberto Villa, and Mark A. Anastasio. "Mining the Manifolds of Deep Generative Models for Multiple Data-Consistent Solutions of Ill-Posed Tomographic Imaging Problems". In: *arXiv*. 2022 (cit. on pp. 63, 72, 89, 91).

[Bissonnette, 2012] Jean-Pierre Bissonnette, Peter A Balter, Lei Dong, Katja M Langen, D Michael Lovelock, Moyed Miften, Douglas J Moseley, Jean Pouliot, Jan-Jakob Sonke, and Sua Yoo. "Quality assurance for image-guided radiation therapy utilizing CT-based technologies: a report of the AAPM TG-179". In: *Medical physics* 39.4 (2012), pp. 1946–1963 (cit. on p. 30).

[Bissonnette, 2008] Jean-Pierre Bissonnette, David J Moseley, and David A Jaffray. "A quality assurance program for image quality of cone-beam CT guidance in radiation therapy". In: *Medical Physics* 35.5 (2008), pp. 1807–1815 (cit. on pp. 17, 29).

[Blundell, 2015] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. "Weight uncertainty in neural networks". In: *International conference on machine learning*. 2015, pp. 1613–1622.

[Bohoudi, 2017] Omar Bohoudi, Anne-Marie E Bruynzeel, Suresh Senan, et al. "Fast and robust online adaptive planning in stereotactic MR-guided adaptive radiation therapy (SMART) for pancreatic cancer". In: *Radiotherapy and Oncology* 125.3 (2017), pp. 439–444 (cit. on pp. 25, 30).

[Bora, 2017] Ashish Bora, Ajil Jalal, Eric Price, and Alexandros G. Dimakis. "Compressed Sensing Using Generative Models". In: *ICML*. 2017 (cit. on pp. 50, 53, 55, 57–59, 66, 71, 83, 89).

[Brock, 2019] Andrew Brock, Jeff Donahue, and Karen Simonyan. "Large scale GAN training for high fidelity natural image synthesis". In: *International Conference on Learning Representations (ICLR)*. 2019 (cit. on p. 60).

[Brouwer, 2015] Charlotte L Brouwer, Roel JHM Steenbakkers, Jean Bourhis, Wilfried Budach, Cai Grau, Vincent Grégoire, Marcel Van Herk, Anne Lee, Philippe Maingon, Chris Nutting, et al. "CT-based delineation of organs at risk in the head and neck region: DAHANCA, EORTC, GORTEC, HKNPCSG, NCIC CTG, NCRI, NRG Oncology and TROG consensus guidelines". In: *Radiotherapy and Oncology* 117.1 (2015), pp. 83–90 (cit. on p. 11).

[Browne, 1996] Jolyon Browne and AB De Pierro. "A row-action alternative to the EM algorithm for maximizing likelihood in emission tomography". In: *IEEE transactions on medical imaging* 15.5 (1996), pp. 687–699 (cit. on p. 48).

[Buatti, 2024] Jacob S Buatti, Alexandre Cafaro, Sruthi Sivabhaskar, Kristen Duke, Nikos Papanikolaou, Neil Kirby, and Nikos Paragios. "2008: A Generative Adversarial Network for Radiotherapy Dose Predictions of Head and Neck Cancers". In: *Radiotherapy and Oncology* 194 (2024), S3618–S3620.

[Bushberg, 2011] Jerrold T. Bushberg, J. Anthony Seibert, Edwin M. Leidholdt Jr, and John M. Boone. *The Essential Physics of Medical Imaging*. 3rd. Lippincott Williams & Wilkins, 2011 (cit. on p. 40).

[Buvat, 2006] Irène Buvat. "Reconstruction tomographique". In: *Cours de Master de Physique Médicale-Université Paris Sud (Orsay)* (2006) (cit. on pp. 37, 41, 44, 45, 48, 49).

[Cafaro, 2023a] A Cafaro, Q Spinat, A Leroy, P Maury, G Beldjoudi, C Robert, E Deutsch, V Grégoire, N Paragios, and V Lepetit. "OC-0443 Full 3D CT reconstruction from partial bi-planar projections using a deep generative model". In: *Radiotherapy and Oncology* (2023). Publisher: Elsevier. Presented at ESTRO 2023.

[Cafaro, 2023b] A Cafaro, Q Spinat, A Leroy, P Maury, G Beldjoudi, C Robert, E Deutsch, V Grégoire, N Paragios, and V Lepetit. "PO-1649 Style-based generative model to reconstruct head and neck 3D CTs". In: *Radiotherapy and Oncology* (2023). Publisher: Elsevier. Presented at ESTRO 2023 (cit. on p. 3).

[Cafaro, 2023c] A Cafaro, Q Spinat, A Leroy, P Maury, A Munoz, et al. "X2Vision: 3D CT Reconstruction from Biplanar X-Rays with Deep Structure Prior". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*. Ed. by Hayit Greenspan, Anant Madabhushi, Parvin Mousavi, Septimiu Salcudean, James Duncan, Tanveer Syeda-Mahmood, and Russell Taylor. Lecture Notes in Computer Science. Cham: Springer Nature Switzerland, 2023, pp. 699–709 (cit. on p. 3).

[Cafaro, 2024a] Alexandre Cafaro, Reuben Dorent, Nazim Haouchine, Vincent Lepetit, Nikos Paragios, William M. Wells III, and Sarah Frisken. "Two Projections Suffice for Cerebral Vascular Reconstruction". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*. Ed. by Marius George Linguraru, Qi Dou, Aasa Feragen, Stamatia Giannarou, Ben Glocker, Karim Lekadir, and Julia A. Schnabel. Cham: Springer Nature Switzerland, 2024, pp. 722–731 (cit. on pp. 4, 165).

[Cafaro, 2024b] Alexandre Cafaro, Amaury Leroy, Guillaume Beldjoudi, Pauline Maury, Charlotte Robert, Eric Deutsch, Vincent Grégoire, Vincent Lepetit, and Nikos Paragios. "XSynthMorph: Generative-Guided Deformation for Unsupervised Ill-Posed Volumetric Recovery". In: *International Workshop on Biomedical Image Registration*. Springer. 2024, pp. 19–33 (cit. on p. 4).

[Cafaro, 2024c] Alexandre Cafaro, Amaury Leroy, Guillaume Beldjoudy, Pauline Maury, Alexandre Munoz, Charlotte Robert, Vincent Lepetit, Nikos Paragios, Vincent Grégoire, and Eric Deutsch. "829: 3D CT Reconstruction from biplanar projections with integration of planning CT". In:

Radiotherapy and Oncology 194 (2024). Publisher: Elsevier. Presented at ESTRO 2024, S3807–S3810 (cit. on p. 4).

[Cafaro, 2024d] Alexandre Cafaro, Amandine Ruffier, Gabriele Bielinyte, Youlia Kirova, Séverine Racadot, Mohamed Benchalal, Jean-Baptiste Clavier, Claire Charra-Brunaud, Marie-Eve Chand-Fouche, Delphine Argo-Leignel, et al. "Abstract PO5-21-03: Cosmetic assessment in the UNICANCER HypoG-01 trial: a deep learning approach". In: *Cancer Research* 84.9_Supplement (2024), PO5–21.

[Cafaro, 2024e] Alexandre Cafaro, Quentin Spinat, Eric Deutsch, Vincent Gregoire, and Nikos Paragios. *3d reconstruction from a limited number of 2d projections*. US Patent App. 18/634,076. Oct. 2024 (cit. on p. 4).

[Cafaro, 2023d] Alexandre Cafaro, Quentin Spinat, Amaury Leroy, Pauline Maury, Alexandre Munoz, et al. "X2Vision: 3D CT Reconstruction from Biplanar X-Rays with Deep Structure Prior". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2023 (cit. on pp. 110, 111, 120, 172).

[Candès, 2006] Emmanuel J Candès, Justin Romberg, and Terence Tao. "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information". In: *IEEE Transactions on Information Theory* 52.2 (2006), pp. 489–509 (cit. on p. 52).

[Capelle, 2012] Lisa Capelle, Maureen Mackenzie, Carrie Field, et al. "Adaptive radiotherapy using helical tomotherapy for head and neck cancer in definitive and postoperative settings: Initial results". In: *Clinical Oncology* 24.3 (2012), pp. 208–215 (cit. on p. 22).

[Casamitjana, 2021] Alex Casamitjana, Mikel Mancini, and Juan Eugenio Iglesias. "Synth-by-Reg (SBR): Contrastive learning for synthesis-based registration of paired images". In: *Simulation and Synthesis in Medical Imaging*. Springer, 2021, pp. 65–74 (cit. on p. 104).

[Castadot, 2010] Pierre Castadot, James A Lee, Xavier Geets, and Vincent Grégoire. "Adaptive radiotherapy of head and neck cancer". In: *Seminars in Radiation Oncology* 20.2 (2010), pp. 84–93 (cit. on p. 22).

[Chan, 2024] Yan Chi Ivy Chan, Minglun Li, Adrian Thummerer, Katia Parodi, Claus Belka, Christopher Kurz, and Guillaume Landry. "Minimum imaging dose for deep learning-based pelvic synthetic computed tomography generation from cone beam images". In: *Physics and Imaging in Radiation Oncology* 30 (2024), p. 100569 (cit. on pp. 31, 32).

[Charters, 2022] John A Charters, Pascal Bertram, and James M Lamb. "Offline generator for digitally reconstructed radiographs of a commercial stereoscopic radiotherapy image-guidance system". In: *Journal of Applied Clinical Medical Physics* 23.3 (2022), e13492 (cit. on pp. 138–141).

[Chen, 2008a] Guang-Hong Chen, Jie Tang, and Shuai Leng. "Prior image constrained compressed sensing (PICCS): a method to accurately reconstruct dynamic CT images from highly undersampled projection data sets". In: *Medical physics* 35.2 (2008), pp. 660–663 (cit. on p. 101).

[Chen, 2008b] Guang-Hong Chen, Jing Tang, and Shuai Leng. "Prior image constrained compressed sensing (PICCS): a method to accurately reconstruct dynamic CT images from highly undersampled projection data sets". In: *Medical physics* 35.2 (2008), pp. 660–663.

[Chen, 2018] Yiqun Chen, Ying Tai, Xiaoming Liu, Chunhua Shen, and Jian Yang. "FSRNet: End-to-end learning face super-resolution with facial priors". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 2492–2501 (cit. on p. 65).

[Chetty, 2019] Indrin J Chetty and Mihaela Rosu-Bubulac. "Deformable registration for dose accumulation". In: *Seminars in Radiation Oncology* 29.3 (2019), pp. 198–208 (cit. on p. 28).

[Chung, 2022] Hyungjin Chung and Jong Chul Ye. "Score-based diffusion models for accelerated MRI". In: *Medical image analysis* 80 (2022), p. 102479 (cit. on p. 71).

[Çiçek, 2016] Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. "3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation". In: *MICCAI*. 2016.

[Cipra, 1987] Barry A Cipra. "An introduction to the Ising model". In: *The American Mathematical Monthly* 94.10 (1987), pp. 937–959 (cit. on pp. 166, 169).

[Clinic, 2024] Cleveland Clinic. *Head and Neck Cancer*. 2024 (cit. on p. 10).

[Copeland, 2010] Andrew D Copeland, Rami S Mangoubi, Mukund N Desai, Sanjoy K Mitter, and Adel M Malek. "Spatio-temporal data fusion for 3D+ T image reconstruction in cerebral angiography". In: *IEEE transactions on medical imaging* 29.6 (2010), pp. 1238–1251 (cit. on p. 166).

[Corona-Figueroa, 2022] Abril Corona-Figueroa, Jonathan Frawley, Sam Bond-Taylor, Sarath Bethapudi, Hubert PH Shum, and Chris G. Willcocks. "Mednerf: Medical Neural Radiance Fields for Reconstructing 3D-Aware Ct-Projections from a Single X-Ray". In: *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. 2022.

[Curie, 1898] Marie Curie and Pierre Curie. "On a new radioactive substance contained in pitchblende". In: *Comptes Rendus de l'Académie des Sciences* 127 (1898), pp. 1215–1217 (cit. on p. 8).

[Cuturi, 2013] Marco Cuturi. "Sinkhorn distances: Lightspeed computation of optimal transport". In: *Advances in neural information processing systems*. 2013, pp. 2292–2300 (cit. on p. 127).

[Daras, 2022] Giannis Daras, Yuval Dagan, Alexandros G Dimakis, and Constantinos Daskalakis. "Score-guided intermediate layer optimization: Fast Langevin mixing for inverse problems". In: *arXiv preprint arXiv:2206.09104* (2022) (cit. on p. 71).

[Daras, 2021] Giannis Daras, Joseph Dean, Ajil Jalal, and Alexandros G. Dimakis. "Intermediate Layer Optimization for Inverse Problems Using Deep Generative Models". In: *arXiv*. 2021.

[Dinh, 2014] Laurent Dinh, David Krueger, and Yoshua Bengio. "Nice: Non-linear independent components estimation". In: *arXiv preprint arXiv:1410.8516* (2014) (cit. on p. 56).

[Dong, 2023] Guoya Dong, Jingjing Dai, Na Li, Chulong Zhang, Wenfeng He, Lin Liu, Yinping Chan, Yunhui Li, Yaoqin Xie, and Xiaokun Liang. "2D/3D Non-Rigid Image Registration via Two Orthogonal X-Ray Projection Images for Lung Tumor Tracking". In: *Bioengineering* 10.2 (2023) (cit. on pp. 108, 110, 113, 114, 120, 121).

[Dong, 2022] Guoya Dong, Chenglong Zhang, Lei Deng, Yulin Zhu, Jingjing Dai, Liming Song, Ruoyan Meng, Tianye Niu, Xiaokun Liang, and Yaoqin Xie. "A deep unsupervised learning framework for the 4D CBCT artifact correction". In: *Physics in Medicine & Biology* 67.5 (2022), p. 055012 (cit. on p. 31).

[Donoho, 2006] David L Donoho. "Compressed sensing". In: *IEEE Transactions on information theory* 52.4 (2006), pp. 1289–1306 (cit. on p. 52).

[DotEagle, 2023] DotEagle. *Filtered Back-Projection reconstructions of Shepp-Logan phantom with different number of projections*. Creative Commons Attribution-Share Alike 4.0 International license. Accessed: November-2023. 2023 (cit. on p. 50).

[Dworzak, 2010] Jalda Dworzak, Hans Lamecker, Jens Von Berg, Tobias Klinder, Cristian Lorenz, Dagmar Kainmüller, Heiko Seim, Hans-Christian Hege, and Stefan Zachow. "3D Reconstruction of the Human Rib Cage from 2D Projection Images Using a Statistical Shape Model". In: *International journal of computer assisted radiology and surgery* 5.2 (2010).

[Edmund, 2017] Jens M Edmund and Tufve Nyholm. "A review of substitute CT generation for MRI-only radiation therapy". In: *Radiation Oncology* 12.1 (2017), p. 28 (cit. on pp. 29, 31).

[Elekta, 2023] Elekta. *Versa HD: Advanced Radiation Therapy System*. [Online; accessed November-2023]. 2023 (cit. on p. 82).

[Elekta, 2024] Elekta. *Elekta Unity: MR-Linac System for Radiation Therapy*. https://www.elekta.com/products/radiation-therapy/unity/. Accessed: 2024. 2024 (cit. on pp. 19, 25).

[Ellis, 2022] Sam Ellis, Octavio E. Martinez Manzanera, Vasileios Baltatzis, Ibrahim Nawaz, Arjun Nair, Loıc Le Folgoc, Sujal Desai, Ben Glocker, and Julia A. Schnabel. "Evaluation of 3D GANs for Lung Tissue Modelling in Pulmonary CT". In: *arXiv*. 2022 (cit. on p. 77).

[Esteban, 2019] Javier Esteban, Matthias Grimm, Mathias Unberath, Guillaume Zahnd, and Nassir Navab. "Towards Fully Automatic X-Ray to CT Registration". In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2019*. Springer, 2019, pp. 631–639 (cit. on p. 107).

[Fan, 2019] Jing Fan, Jun Wang, Zhen Chen, et al. "Automatic treatment planning based on three-dimensional dose distribution predicted from deep learning technique". In: *Medical Physics* 46.1 (2019), pp. 370–381 (cit. on p. 28).

[Feldkamp, 1984a] Lee A Feldkamp, Lloyd C Davis, and James W Kress. "Practical cone-beam algorithm". In: *Josa a* 1.6 (1984), pp. 612–619 (cit. on pp. 44, 85, 111, 123).

[Feldkamp, 1984b] Lee A. Feldkamp, L. C. Davis, and James W. Kress. "Practical cone-beam algorithm". In: *Journal of The Optical Society of America A-optics Image Science and Vision* (1984).

[Ferrante, 2017] Enzo Ferrante and Nikos Paragios. "Slice-to-volume medical image registration: A survey". In: *Medical image analysis* 39 (2017), pp. 101–123 (cit. on p. 105).

[Fischer-Valuck, 2017] Benjamin W Fischer-Valuck, Laura Henke, Olga Green, et al. "Two-and-a-half-year clinical experience with the world's first magnetic resonance image guided radiation therapy system". In: *Advances in Radiation Oncology* 2.3 (2017), pp. 485–493.

[Flach, 2014] Barbara Flach, Marcus Brehm, Stefan Sawall, and Marc Kachelrieß. "Deformable 3D–2D Registration for CT and Its Application to Low Dose Tomographic Fluoroscopy". In: *Physics in Medicine & Biology* 59.24 (2014) (cit. on p. 111).

[Foote, 2019] Markus D. Foote, Blake E. Zimmerman, Amit Sawant, and Sarang C. Joshi. "Real-Time 2D-3D Deformable Registration with Deep Learning and Application to Lung Radiotherapy Targeting". In: 2019 (cit. on pp. 111, 112).

[Fortunati, 2014] V. Fortunati, R. F. Verhaart, F. Angeloni, and et al. "Feasibility of multimodal deformable registration for head and neck tumor treatment planning". In: *International Journal of Radiation Oncology* (2014) (cit. on p. 105).

[Frisken, 2022] Sarah Frisken, Nazim Haouchine, Rose Du, and Alexandra J Golby. "Using temporal and structural data to reconstruct 3D cerebral vasculature from a pair of 2D digital subtraction angiography sequences". In: *Computerized Medical Imaging and Graphics* 99 (2022), p. 102076 (cit. on p. 166).

[Frisken, 2024] SF Frisken, N Haouchine, DD Chlorogiannis, V Gopalakrishnan, A Cafaro, WT Wells, AJ Golby, and R Du. "VESCL: an open source 2D vessel contouring library". In: *International Journal of Computer Assisted Radiology and Surgery* (2024), pp. 1–10.

[Frysch, 2021] Robert Frysch, Tim Pfeiffer, and Georg Rose. "A Novel Approach to 2D/3D Registration of X-Ray Images Using Grangeat's Relation". In: *Medical Image Analysis*. 2021.

[Gao, 2020] Cong Gao, Robert B. Grupp, Mathias Unberath, Russell H. Taylor, and Mehran Armand. "Fiducial-Free 2D/3D Registration of the Proximal Femur for Robot-Assisted Femoroplasty". In: *Medical Imaging 2020: Image-Guided Procedures, Robotic Interventions, and Modeling*. 2020.

[Gao, 2021] Liugang Gao, Kai Xie, Xiaojin Wu, Zhengda Lu, Chunying Li, Jiawei Sun, Tao Lin, Jianfeng Sui, and Xinye Ni. "Generating synthetic CT from low-dose cone-beam CT by using generative adversarial networks for adaptive radiotherapy". In: *Radiation Oncology* 16 (2021), pp. 1–16 (cit. on p. 31).

[Glocker, 2008] Ben Glocker, Nikos Komodakis, Georgios Tziritas, Nassir Navab, and Nikos Paragios. "Dense Image Registration through MRFs and Efficient Linear Programming". In: *Medical Image Analysis*. 2008 (cit. on p. 82).

[Goodfellow, 2014] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative adversarial nets". In: *Advances in neural information processing systems* 27 (2014) (cit. on pp. 55, 56).

[Goodfellow, 2020] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative Adversarial Networks". In: *Communications of the ACM* 63.11 (2020) (cit. on p. 80).

[Gopalakrishnan, 2024] Vivek Gopalakrishnan, Neel Dey, and Polina Golland. "Intraoperative 2d/3d image registration via differentiable x-ray rendering". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 11662–11672 (cit. on pp. 107, 108).

[Gordon, 1970] Richard Gordon, Robert Bender, and Gabor T Herman. "Algebraic reconstruction techniques (ART) for three-dimensional electron microscopy and X-ray photography". In: *Journal of theoretical Biology* 29.3 (1970), pp. 471–481 (cit. on p. 47).

[Graham, 2007] SA Graham, DJ Moseley, JH Siewerdsen, and DA Jaffray. "Compensation of scatter in cone-beam CT: effect on image noise, contrast, and cupping artifacts". In: *Medical physics* 34.5 (2007), pp. 2007–2016 (cit. on p. 44).

[Graves, 2011] Alex Graves. "Practical variational inference for neural networks". In: *Advances in neural information processing systems*. 2011, pp. 2348–2356 (cit. on p. 69).

[Grégoire, 2003] Vincent Grégoire, Peter Levendag, Kian K Ang, Jacques Bernier, Marijel Braaksma, Volker Budach, Cliff Chao, Emmanuel Coche, Jay S Cooper, Guy Cosnard, et al. "CT-based delineation of lymph node levels and related CTVs in the node-negative neck: DAHANCA, EORTC, GORTEC, NCIC, RTOG consensus guidelines". In: *Radiotherapy and oncology* 69.3 (2003), pp. 227–236 (cit. on p. 11).

[Grossberg, 2020] A. Grossberg, H. Elhalawani, A. Mohamed, S. Mulder, B. Williams, et al. "Anderson Cancer Center Head and Neck Quantitative Imaging Working Group. HNSCC". In: (2020) (cit. on p. 82).

[Hammernik, 2018] Kerstin Hammernik, Teresa Klatzer, Erich Kobler, Michael P Recht, Daniel K Sodickson, Thomas Pock, and Florian Knoll. "Learning a variational network for reconstruction of accelerated MRI data". In:

Magnetic resonance in medicine 79.6 (2018), pp. 3055–3071 (cit. on p. 55).

[Hammersmith Hospital London,] Hammersmith Hospital London. *IXI Dataset: Brain Development*. https://brain-development.org/ixi-dataset/ (cit. on pp. 170, 171).

[Hansen, 2018] David C Hansen, Guillaume Landry, Florian Kamp, Minglun Li, Claus Belka, Katia Parodi, and Christopher Kurz. "ScatterNet: a convolutional neural network for cone-beam CT intensity correction". In: *Medical physics* 45.11 (2018), pp. 4916–4926 (cit. on pp. 31, 145).

[Hansen, 2006] Per Christian Hansen, James G Nagy, and Dianne P O'leary. *Deblurring images: matrices, spectra, and filtering*. SIAM, 2006 (cit. on pp. 47, 48).

[Haouchine, 2021] Nazim Haouchine, Parikshit Juvekar, Xin Xiong, Jie Luo, Tina Kapur, Rose Du, Alexandra Golby, and Sarah Frisken. "Estimation of high framerate digital subtraction angiography sequences at low radiation dose". In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VI 24*. Springer. 2021, pp. 171–180 (cit. on pp. 165, 166).

[Heijkoop, 2014] Sander T Heijkoop, Thomas R Langerak, Sjoerd Quint, et al. "Clinical implementation of an online adaptive plan-of-the-day protocol for nonrigid motion management in locally advanced cervical cancer IMRT". In: *International Journal of Radiation Oncology* Biology* Physics* 90.3 (2014), pp. 673–679 (cit. on p. 22).

[Hendee, 1986] William R Hendee and F Marc Edwards. "ALARA and an integrated approach to radiation protection". In: *Seminars in Nuclear Medicine*. Vol. 16. 2. Elsevier. 1986, pp. 142–150 (cit. on p. 31).

[Henzler, 2018] Phlipp Henzler, Volker Rasche, Timo Ropinski, and Tobias Ritschel. "Single-Image Tomography: 3D Volumes from 2D Cranial X-Rays". In: *Computer Graphics Forum*. 2018 (cit. on pp. 73, 108, 110, 111).

[Herman, 2009] Gabor T. Herman. *Fundamentals of Computerized Tomography: Image Reconstruction from Projections*. Springer Science & Business Media, 2009 (cit. on pp. 41, 43–45, 47).

[Herrmann, 2015] H. Herrmann, Y. Seppenwoolde, D. Georg, and J. Widder. "Image guidance: past and future of radiotherapy". In: *Department of Radiotherapy, Medical University of Vienna, Vienna, Austria* (2015) (cit. on p. 21).

[Heukelom, 2019] Jolien Heukelom and Clifton David Fuller. "Head and neck cancer adaptive radiation therapy (ART): conceptual considerations for the informed clinician". In: *Seminars in radiation oncology*. Vol. 29. 3. Elsevier. 2019, pp. 258–273 (cit. on pp. 23–27).

[Heusel, 2017] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. "Gans Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium". In: *NeurIPS*. 2017 (cit. on p. 83).

[Hinton, 1993] Geoffrey E Hinton and Drew Van Camp. "Keeping the neural networks simple by minimizing the description length of the weights". In: *Proceedings of the sixth annual conference on Computational learning theory*. 1993, pp. 5–13 (cit. on p. 69).

[Hoffman, 2017] Matthew D Hoffman and Matthew J Johnson. "Learning deep latent Gaussian models with Markov chain Monte Carlo". In: *Proceedings of the 34th International Conference on Machine Learning (ICML)*. 2017, pp. 1510–1519 (cit. on p. 60).

[Holden, 2022] Matthew Holden, Marcelo Pereyra, and Konstantinos C Zygalakis. "Bayesian imaging with data-driven priors encoded by neural net-

works". In: *SIAM Journal on Imaging Sciences* 15.2 (2022), pp. 892–924 (cit. on p. 71).

[Hong, 2021] Sungmin Hong, Razvan Marinescu, Adrian V. Dalca, Anna K. Bonkhoff, Martin Bretzner, Natalia S. Rost, and Polina Golland. "3D-Stylegan: A Style-Based Generative Adversarial Network for Generative Modeling of Three-Dimensional Medical Images". In: *Deep Generative Models, and Data Augmentation, Labelling, and Imperfections*. Springer, 2021 (cit. on pp. 77, 78, 80, 83, 116).

[Huang, 2005] Xishi Huang, Nicholas A Hill, Jing Ren, Gerard Guiraudon, Derek Boughner, and Terry M Peters. "Dynamic 3D ultrasound and MR image registration of the beating heart". In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2005: 8th International Conference, Palm Springs, CA, USA, October 26-29, 2005, Proceedings, Part II 8*. Springer. 2005, pp. 171–178 (cit. on p. 106).

[Huang, 2017] Xun Huang and Serge Belongie. "Arbitrary style transfer in real-time with adaptive instance normalization". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 1501–1510 (cit. on p. 62).

[Hubbell, 1995] J. H. Hubbell. "Tables of X-Ray Mass Attenuation Coefficients 1 keV to 20 MeV for Elements Z= 1 to 92 and 48 Additional Substance of Dosimetric Interest". In: *NISTIR 5632* (1995) (cit. on pp. 40, 81).

[Hudson, 1994] H Malcolm Hudson and Richard S Larkin. "Accelerated image reconstruction using ordered subsets of projection data". In: *IEEE transactions on medical imaging* 13.4 (1994), pp. 601–609 (cit. on p. 48).

[Hussein, 2018] Mohamed Hussein, Ben JM Heijmen, Dirk Verellen, et al. "Automation in intensity modulated radiotherapy treatment planning: A review of recent innovations". In: *British Journal of Radiology* 91.1091 (2018), p. 20180270 (cit. on pp. 22, 25, 28, 30).

[Isensee, 2021] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation". In: *Nature methods* 18.2 (2021), pp. 203–211 (cit. on p. 167).

[Isola, 2017] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. "Image-to-image translation with conditional adversarial networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 1125–1134 (cit. on p. 104).

[Jaderberg, 2015] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. "Spatial Transformer Networks". In: *NeurIPS*. 2015 (cit. on p. 115).

[Jaganathan, 2023] Srikrishna Jaganathan, Maximilian Kukla, Jian Wang, Karthik Shetty, and Andreas Maier. "Self-Supervised 2D/3D Registration for X-Ray to CT Image Fusion". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 2023, pp. 2788–2798 (cit. on p. 107).

[Jalal, 2021a] Ajil Jalal, Marius Arvinte, Giannis Daras, Eric Price, Alexandros G Dimakis, and Jon Tamir. "Robust compressed sensing mri with deep generative priors". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 14938–14954 (cit. on p. 71).

[Jalal, 2021b] Ajil Jalal, Marius Arvinte, Giannis Daras, Eric Price, Alexandros G. Dimakis, and Jon Tamir. "Robust Compressed Sensing MRI with Deep Generative Priors". In: *NeurIPS*. 2021.

[Jiang, 2022] Yixiang Jiang. "MFCT-GAN: Multi-Information Network to Reconstruct CT Volumes for Security Screening". In: *Journal of Intelligent Manufacturing and Special Equipment* (2022) (cit. on pp. 73, 84, 108, 110, 111).

[Jin, 2010] Jian-Yue Jin, Lei Ren, Qiang Liu, et al. "Combining scatter reduction and correction to improve image quality in cone-beam computed tomography (CBCT)". In: *Medical Physics* 37.11 (2010), pp. 5634–5644 (cit. on p. 31).

[Jin, 2006] Jian-Yue Jin, Samuel Ryu, Kathleen Faber, Tom Mikkelsen, Qing Chen, Shidong Li, and Benjamin Movsas. "2D/3D image fusion for accurate target localization and evaluation of a mask based stereotactic system in fractionated stereotactic radiotherapy of cranial lesions". In: *Medical physics* 33.12 (2006), pp. 4557–4566 (cit. on p. 135).

[Jin, 2008] Jian-Yue Jin, Fang-Fang Yin, Stephen E Tenn, Paul M Medin, and Timothy D Solberg. "Use of the BrainLAB ExacTrac X-Ray 6D system in image-guided radiotherapy". In: *Medical Dosimetry* 33.2 (2008), pp. 124–134 (cit. on pp. 17, 134, 135, 137).

[Jin, 2017] Kyong Hwan Jin, Michael T McCann, Emmanuel Froustey, and Michael Unser. "Deep convolutional neural network for inverse problems in imaging". In: *IEEE transactions on image processing* 26.9 (2017), pp. 4509–4522 (cit. on p. 53).

[Johnson, 2016] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. "Perceptual losses for real-time style transfer and super-resolution". In: *European conference on computer vision*. 2016 (cit. on pp. 68, 79, 117).

[Kaipio, 2006] Jari Kaipio and Erkki Somersalo. *Statistical and computational inverse problems*. Vol. 160. Springer Science & Business Media, 2006 (cit. on pp. 50, 51).

[Kak, 2001] Avinash C Kak and Malcolm Slaney. *Principles of computerized tomographic imaging*. SIAM, 2001 (cit. on pp. 36, 40, 42–44).

[Karras, 2020a] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. "Training Generative Adversarial Networks with Limited Data". In: *NeurIPS*. 2020 (cit. on p. 80).

[Karras, 2019] Tero Karras, Samuli Laine, and Timo Aila. "A style-based generator architecture for generative adversarial networks". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 4401–4410 (cit. on pp. 60–62, 64, 70, 77, 102).

[Karras, 2020b] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. "Analyzing and Improving the Image Quality of Stylegan". In: *CVPR*. 2020 (cit. on pp. 60, 63, 65, 77, 80, 82, 83).

[Kelkar, 2021a] Varun A Kelkar and Mark A Anastasio. "Compressible latent-space invertible networks for generative model-constrained image reconstruction". In: *IEEE Transactions on Computational Imaging* 7 (2021), pp. 209–223 (cit. on pp. 101–103).

[Kelkar, 2023a] Varun A Kelkar, Rucha Deshpande, Arindam Banerjee, and Mark A Anastasio. "AmbientFlow: Invertible generative models from incomplete, noisy measurements". In: *arXiv preprint arXiv:2309.04856* (2023) (cit. on p. 71).

[Kelkar, 2023b] Varun A Kelkar, Dimitrios S Gotsis, Frank J Brooks, KC Prabhat, Kyle J Myers, Rongping Zeng, and Mark A Anastasio. "Assessing the ability of generative adversarial networks to learn canonical medical image statistics". In: *IEEE transactions on medical imaging* 42.6 (2023), pp. 1799–1808 (cit. on p. 71).

[Kelkar, 2021b] Varun A. Kelkar and Mark Anastasio. "Prior Image-Constrained Reconstruction Using Style-Based Generative Models". In: *ICML*. 2021 (cit. on p. 71).

[Kerbl, 2023] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. "3D Gaussian Splatting for Real-Time Radiance Field

Rendering." In: *ACM Trans. Graph.* 42.4 (2023), pp. 139–1 (cit. on p. 76).

[Ketcha, 2017] M. D. Ketcha, T. De Silva, A. Uneri, M. W. Jacobson, J. Goerres, G. Kleinszig, S. Vogt, J. P. Wolinsky, and J. H. Siewerdsen. "Multi-Stage 3D–2D Registration for Correction of Anatomical Deformation in Image-Guided Spine Surgery". In: *Physics in Medicine & Biology* 62.11 (2017).

[Kinahan, 2020] P. Kinahan, M. Muzi, B. Bialecki, and L. Coombs. *Data from the ACRIN 6685 Trial HNSCC-FDG-PET/CT*. 2020 (cit. on p. 82).

[Kingma, 2013] Diederik P Kingma. "Auto-encoding variational bayes". In: *arXiv preprint arXiv:1312.6114* (2013) (cit. on p. 56).

[Kingma, 2014] Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization". In: *arXiv*. 2014 (cit. on pp. 66, 171).

[Kingma, 2018] Durk P Kingma and Prafulla Dhariwal. "Glow: Generative flow with invertible 1x1 convolutions". In: *Advances in neural information processing systems* 31 (2018).

[Kranen, 2013] Sander van Kranen, Andrea Mencarelli, Suzanne van Beek, et al. "Adaptive radiotherapy with an average anatomy model: Evaluation and quantification of residual deformations in head and neck cancer patients". In: *Radiotherapy and Oncology* 109.3 (2013), pp. 463–468 (cit. on p. 22).

[Kruse, 2021] Jakob Kruse, Ching-An Wu, Pablo Marquez-Neila, Jonas Kohler, Bernhard Schölkopf, Jan Peters, and Wouter Lueks. "HINT: Hierarchical Invertible Neural Transport for General High-Dimensional Bayesian Inference". In: *Proceedings of the 38th International Conference on Machine Learning (ICML)*. 2021, pp. 5658–5667 (cit. on p. 60).

[Kupelian, 2014] Patrick Kupelian and Jan-Jakob Sonke. "Magnetic resonance-guided adaptive radiotherapy: A solution to the future". In: *Seminars in Radiation Oncology* 24.3 (2014), pp. 227–232 (cit. on p. 29).

[Kupyn, 2018] Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, and Jiri Matas. "DeblurGAN: Blind motion deblurring using conditional adversarial networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 8183–8192 (cit. on p. 54).

[Kupyn, 2019] Orest Kupyn, Tetiana Martyniuk, Junru Wu, and Zhangyang Wang. "Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 8878–8887 (cit. on p. 54).

[Kurz, 2016] Christopher Kurz, Florian Kamp, Yang-Kyun Park, Christoph Zöllner, Simon Rit, David Hansen, Mark Podesta, Gregory C Sharp, Minglun Li, Michael Reiner, et al. "Investigating deformable image registration and scatter correction for CBCT-based dose calculation in adaptive IMPT". In: *Medical physics* 43.10 (2016), pp. 5635–5646 (cit. on p. 145).

[Kurz, 2019] Christopher Kurz, Matteo Maspero, Mark HF Savenije, Guillaume Landry, Florian Kamp, Marco Pinto, Minglun Li, Katia Parodi, Claus Belka, and Cornelis AT Van den Berg. "CBCT correction using a cycle-consistent generative adversarial network and unpaired training to enable photon and proton dose calculation". In: *Physics in Medicine & Biology* 64.22 (2019), p. 225004 (cit. on p. 31).

[Kwan, 2019] J. Y. Y. Kwan, J. Su, S. H. Huang, L. S. Ghoraie, W. Xu, et al. "Data from Radiomic Biomarkers to Refine Risk Models for Distant Metastasis in Oropharyngeal Carcinoma". In: (2019) (cit. on p. 82).

[Landry, 2019] Guillaume Landry, David Hansen, Florian Kamp, Minglun Li, Ben Hoyle, Jochen Weller, Katia Parodi, Claus Belka, and Christopher

Kurz. "Comparing Unet training with three different datasets to correct CBCT images for prostate radiotherapy dose calculations". In: *Physics in Medicine & Biology* 64.3 (2019), p. 035011 (cit. on p. 31).

[Lange, 2003] Thomas Lange, Sebastian Eulenstein, Michael Hünerbein, and Peter-Michael Schlag. "Vessel-based non-rigid registration of MR/CT and 3D ultrasound for navigation in liver surgery". In: *Computer Aided Surgery* 8.5 (2003), pp. 228–240.

[Lechuga, 2016] Lawrence Lechuga and Georg A Weidlich. "Cone beam CT vs. fan beam CT: a comparison of image quality and dose delivered between two differing CT imaging modalities". In: *Cureus* 8.9 (2016) (cit. on p. 46).

[Lecomte, 2022] François Lecomte, Jean-Louis Dillenseger, and Stéphane Cotin. "CNN-based real-time 2D-3D deformable registration from a single X-ray projection". In: *arXiv preprint arXiv:2212.07692* (2022) (cit. on p. 108).

[LeCun, 1998] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324 (cit. on p. 59).

[Ledig, 2017] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. "Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network". In: *CVPR*. 2017 (cit. on p. 54).

[Lefebvre, 2010] J-L Lefebvre, L Licitra, and E Felip. "Squamous cell carcinoma of the head and neck: EHNS–ESMO–ESTRO Clinical Practice Guidelines for diagnosis, treatment and follow-up". In: *Annals of oncology* 21 (2010), pp. v184–v186 (cit. on p. 11).

[Lemieux, 1994] L Lemieux, R Jagoe, DR Fish, ND Kitchen, and DGT Thomas. "A patient-to-computed-tomography image registration method based on digitally reconstructed radiographs". In: *Medical physics* 21.11 (1994), pp. 1749–1760 (cit. on p. 135).

[Lempitsky, 2018] Victor Lempitsky, Andrea Vedaldi, and Dmitry Ulyanov. "Deep image prior". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE. 2018, pp. 9446–9454.

[Lepetit, 2009] Vincent Lepetit, Francesc Moreno-Noguer, and Pascal Fua. "EP n P: An accurate O (n) solution to the P n P problem". In: *International journal of computer vision* 81 (2009), pp. 155–166 (cit. on p. 140).

[Leroy, 2023a] A Leroy, A Cafaro, G Gessain, A Champagnac, V Grégoire, E Deutsch, V Lepetit, and N Paragios. "StructuRegNet: Structure-Guided Multimodal 2D-3D Registration". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*. Ed. by Hayit Greenspan, Anant Madabhushi, Parvin Mousavi, Septimiu Salcudean, James Duncan, Tanveer Syeda-Mahmood, and Russell Taylor. Lecture Notes in Computer Science. Cham: Springer Nature Switzerland. Presented at MICCAI, 2023, pp. 771–780 (cit. on pp. 104, 106).

[Leroy, 2023b] A Leroy, A Cafaro, V Lepetit, N Paragios, E Deutsch, and V Grégoire. "MO-0714 Statistical comparison between GTV and gold standard contour on AI-based registered histopathology". In: *Radiotherapy and Oncology* (2023). Publisher: Elsevier. Presented at ESTRO 2023.

[Leroy, 2023c] A Leroy, A Cafaro, V Lepetit, N Paragios, E Deutsch, and V Grégoire. "OC-0448 Bridging the gap between radiology and histology through AI-driven registration and reconstruction". In: *Radiotherapy and Oncology* (2023). Publisher: Elsevier. Presented at ESTRO 2023.

[Leroy, 2022] A Leroy, M Lerousseau, T Henry, A Cafaro, N Paragios, V Grégoire, and E Deutsch. "End-to-End Multi-Slice-to-Volume Concurrent Registration and Multimodal Generation". In: *Medical Image Computing and Computer Assisted Intervention - MICCAI 2022*. Cham: Springer Nature Switzerland. Presented at MICCAI, 2022, pp. 152–162.

[Leroy, 2024] Amaury Leroy, Alexandre Cafaro, Nazim Benzerdjeb, Anne Champagnac, Grégoire Gessain, Philippe Gorphe, Daphné Morel, Charlotte Robert, Roger Sun, Philippe Zrounba, et al. "1334: Histology to CT transfer for HNC target volume auto-segmentation with deep-learning diffusion models". In: *Radiotherapy and Oncology* 194 (2024), S3047–S3051.

[Li, 2017] L. Li, S. Xu, and et al. "Co-registration of ex vivo surgical histopathology and in vivo T2-weighted MRI of the prostate via multi-scale spectral embedding representation". In: *Scientific Reports* 7.1 (2017), p. 8717 (cit. on p. 106).

[Li, 2020] Peixin Li, Yuru Pei, Yuke Guo, Gengyu Ma, Tianmin Xu, and Hongbin Zha. "Non-Rigid 2D-3D Registration Using Convolutional Autoencoders". In: 2020 (cit. on pp. 70, 111, 112).

[Li, 2019] Zhen Li, Jinglei Yang, Zheng Liu, Xiaomin Yang, Gwanggil Jeon, and Wei Wu. "Feedback network for image super-resolution". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 3867–3876 (cit. on p. 70).

[Liang, 2019] Xiao Liang, Liyuan Chen, Dan Nguyen, Zhiguo Zhou, Xuejun Gu, Ming Yang, Jing Wang, and Steve Jiang. "Generating synthesized computed tomography (CT) from cone-beam computed tomography (CBCT) using CycleGAN for adaptive radiation therapy". In: *Physics in Medicine & Biology* 64.12 (2019), p. 125002 (cit. on p. 31).

[Lim, 2017] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. "Enhanced deep residual networks for single image super-resolution". In: *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 2017, pp. 136–144 (cit. on p. 54).

[Lin, 2023] Yiqun Lin, Zhongjin Luo, Wei Zhao, and Xiaomeng Li. "Learning Deep Intensity Field for Extremely Sparse-View CBCT Reconstruction". In: *arXiv*. 2023.

[Lin, 2024] Yiqun Lin, Hualiang Wang, Jixiang Chen, and Xiaomeng Li. "Learning 3D Gaussians for Extremely Sparse-View Cone-Beam CT Reconstruction". In: *arXiv preprint arXiv:2407.01090* (2024) (cit. on p. 76).

[Liu, 2015a] Yilin Liu, Fang-Fang Yin, Nan-kuei Chen, Mei-Lan Chu, and Jing Cai. "Four dimensional magnetic resonance imaging with retrospective k-space reordering: a feasibility study". In: *Medical physics* 42.2 (2015), pp. 534–541 (cit. on p. 31).

[Liu, 2023] Yuchen Liu, Wei Wang, Yan Li, et al. "Geometry-consistent adversarial registration model for unsupervised multi-modal medical image registration". In: *IEEE Journal of Biomedical and Health Informatics* 27 (2023), pp. 1091–1100 (cit. on p. 104).

[Liu, 2015b] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. "Deep learning face attributes in the wild". In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 3730–3738 (cit. on p. 59).

[Livyatan, 2003] H. Livyatan, Z. Yaniv, and L. Joskowicz. "Gradient-based 2-D/3-D rigid registration of fluoroscopic X-ray to CT". In: *IEEE Transactions on Medical Imaging* 22.11 (2003), pp. 1395–1406 (cit. on p. 107).

[Lu, 2022] Shaolin Lu, Shibo Li, Yu Wang, Lihai Zhang, Ying Hu, and Bing Li. "Prior Information-Based High-Resolution Tomography Image Re-

construction from a Single Digitally Reconstructed Radiograph". In: *Physics in Medicine & Biology* 67.8 (Apr. 2022) (cit. on pp. 73, 108, 110, 111).

[Lustig, 2007] Michael Lustig, David Donoho, and John M Pauly. "Sparse MRI: The application of compressed sensing for rapid MR imaging". In: *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine* 58.6 (2007), pp. 1182–1195 (cit. on p. 53).

[Lustig, 2008] Michael Lustig, David L Donoho, and John M Pauly. "Compressed sensing MRI". In: *IEEE Signal Processing Magazine* 25.2 (2008), pp. 72–82 (cit. on p. 60).

[Maas, 2023] Kirsten WH Maas, Nicola Pezzotti, Amy JE Vermeer, Danny Ruijters, and Anna Vilanova. "NeRF for 3D Reconstruction from X-ray Angiography: Possibilities and Limitations". In: *VCBM 2023: Eurographics Workshop on Visual Computing for Biology and Medicine*. Eurographics Association. 2023, pp. 29–40 (cit. on p. 166).

[Maken, 2023] Payal Maken and Abhishek Gupta. "2D-to-3D: a review for computational 3D image reconstruction from X-ray images". In: *Archives of Computational Methods in Engineering* 30.1 (2023), pp. 85–114 (cit. on pp. 38, 42).

[Mardani, 2018] Morteza Mardani, Enhao Gong, Joseph Y Cheng, Shreyas S Vasanawala, Greg Zaharchuk, Lei Xing, and John M Pauly. "Deep generative adversarial neural networks for compressive sensing MRI". In: *IEEE transactions on medical imaging* 38.1 (2018), pp. 167–179 (cit. on p. 55).

[Marinescu, 2020] Razvan V. Marinescu, Daniel Moyer, and Polina Golland. "Bayesian Image Reconstruction Using Deep Generative Models". In: *arXiv*. 2020 (cit. on pp. 54, 67, 70, 72, 80, 89).

[Markelj, 2012] Primoz Markelj, Dejan Tomaževič, Bostjan Likar, and Franjo Pernuš. "A review of 3D/2D registration methods for image-guided interventions". In: *Medical image analysis* 16.3 (2012), pp. 642–661 (cit. on p. 106).

[Mencarelli, 2014] A. Mencarelli, S. R. van Kranen, and et al. "Deformable image registration for adaptive radiation therapy of head and neck cancer: accuracy and precision in the presence of tumor changes". In: *International Journal of Radiation Oncology* (2014) (cit. on p. 105).

[Menon, 2020] Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. "Pulse: Self-Supervised Photo Upsampling via Latent Space Exploration of Generative Models". In: *CVPR*. 2020 (cit. on pp. 53, 54, 65, 67–70, 72, 75, 91).

[Mescheder, 2018] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. "Which Training Methods for GANs Do Actually Converge?" In: *ICML*. 2018 (cit. on p. 80).

[Micikevicius, 2017] Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, et al. "Mixed Precision Training". In: *arXiv*. 2017.

[Mildenhall, 2021] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. "NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis". In: *Communications of the ACM* 65.1 (2021) (cit. on pp. 75, 90).

[Miracle, 2009] AC Miracle and SK Mukherji. "Conebeam CT of the head and neck, part 1: physical principles". In: *American Journal of Neuroradiology* 30.6 (2009), pp. 1088–1095 (cit. on p. 45).

[Møller, 2016] Ditte S Møller, Martin I Holt, Markus Alber, et al. "Adaptive radiotherapy for advanced lung cancer ensures target coverage and

decreases lung dose". In: *Radiotherapy and Oncology* 121.1 (2016), pp. 32–38 (cit. on p. 24).

[Navran, 2019] Arash Navran, Wilma Heemsbergen, Tomas Janssen, et al. "The impact of margin reduction on outcome and toxicity in head and neck cancer patients treated with image-guided volumetric modulated arc therapy (VMAT)". In: *Radiotherapy and Oncology* 130 (2019), pp. 25–31 (cit. on pp. 20, 22).

[Nazeri, 2019] K Nazeri. "EdgeConnect: Generative Image Inpainting with Adversarial Edge Learning". In: *arXiv preprint arXiv:1901.00212* (2019) (cit. on p. 54).

[Nguyen, 2019] Dan Nguyen, Xun Jia, David Sher, Mu-Han Lin, Zohaib Iqbal, Hui Liu, and Steve Jiang. "3D radiotherapy dose prediction on head and neck cancer patients with a hierarchically densely connected U-net deep learning architecture". In: *Physics in medicine & Biology* 64.6 (2019), p. 065020 (cit. on p. 28).

[Nguyen-Phuoc, 2019] Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, and Yong-Liang Yang. "Hologan: Unsupervised learning of 3d representations from natural images". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 7588–7597 (cit. on p. 70).

[Niu, 2010] Tianye Niu, Mingshan Sun, Josh Star-Lack, Hewei Gao, Qiyong Fan, and Lei Zhu. "Shading correction for on-board cone-beam CT in radiation therapy using planning MDCT images". In: *Medical physics* 37.10 (2010), pp. 5395–5406 (cit. on p. 145).

[Ohnishi, 2016] T. Ohnishi, M. Yamamoto, and et al. "Deformable image registration between pathological images and MR image via an optical macro image". In: *Pathology Research and Practice* 212.10 (2016), pp. 927–936 (cit. on p. 106).

[Oktay, 2018] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. "Attention u-net: Learning where to look for the pancreas". In: *arXiv preprint arXiv:1804.03999* (2018) (cit. on p. 146).

[Olberg, 2018] Sebastian Olberg, Olga Green, Bin Cai, et al. "Optimization of treatment planning workflow and tumor coverage during daily adaptive magnetic resonance image guided radiation therapy (MR-IGRT) of pancreatic cancer". In: *Radiation Oncology* 13.1 (2018), p. 51 (cit. on p. 28).

[Ongie, 2020] Gregory Ongie, Ajil Jalal, Christopher A Metzler, Richard G Baraniuk, Alexandros G Dimakis, and Rebecca Willett. "Deep learning techniques for inverse problems in imaging". In: *IEEE Journal on Selected Areas in Information Theory* 1.1 (2020), pp. 39–56 (cit. on pp. 53, 54).

[Paetzold, 2019] Johannes C Paetzold, Suprosanna Shit, Ivan Ezhov, Giles Tetteh, Ali Ertürk, Helmholtz Zentrum Munich, and Bjoern Menze. "clDice—A novel connectivity-preserving loss function for vessel segmentation". In: *Medical Imaging Meets NeurIPS 2019 Workshop*. 2019 (cit. on p. 172).

[Paganelli, 2018] Claudia Paganelli, Giulia Meschini, Sara Molinelli, et al. "Patient-specific validation of deformable image registration in radiation therapy: Overview and caveats". In: *Medical Physics* 45.10 (2018) (cit. on pp. 28, 30).

[Park, 2020] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. "Contrastive learning for unpaired image-to-image translation". In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*. Springer. 2020, pp. 319–345 (cit. on p. 31).

[Paszke, 2019] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. "PyTorch: An Imperative Style, High-Performance Deep Learning Library". In: *NeurIPS*. 2019 (cit. on pp. 82, 171).

[Pei, 2017] Yuru Pei, Yungeng Zhang, Haifang Qin, Gengyu Ma, Yuke Guo, Tianmin Xu, and Hongbin Zha. "Non-Rigid Craniofacial 2D-3D Registration using CNN-Based Regression". In: *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI*. 2017 (cit. on pp. 111, 112).

[Peng, 2021] Cheng Peng, Haofu Liao, Gina Wong, Jiebo Luo, S. Kevin Zhou, and Rama Chellappa. "XraySyn: Realistic View Synthesis from a Single Radiograph through CT Priors". In: *AAAI*. 2021 (cit. on pp. 40, 81, 129, 145).

[Penney, 2004] Graeme P Penney, Jane M Blackall, MS Hamady, T Sabharwal, A Adam, and David J Hawkes. "Registration of freehand 3D ultrasound and magnetic resonance liver images". In: *Medical image analysis* 8.1 (2004), pp. 81–91 (cit. on p. 106).

[Pielawski, 2020] Nicolas Pielawski, Elisabeth Wetzer, Johan Öfverstedt, et al. "CoMIR: Contrastive multimodal image representation for registration". In: *Advances in Neural Information Processing Systems*. 2020 (cit. on p. 104).

[Pinaya, 2022] Walter HL Pinaya, Petru-Daniel Tudosiu, Jessica Dafflon, Pedro F Da Costa, Virginia Fernandez, Parashkev Nachev, Sebastien Ourselin, and M Jorge Cardoso. "Brain imaging generation with latent diffusion models". In: *MICCAI Workshop on Deep Generative Models*. Springer. 2022.

[Poludniowski, 2009] Gavin Poludniowski, Guillaume Landry, François Deblois, Philip M Evans, and Frank Verhaegen. "SpekCalc: a program to calculate photon spectra from tungsten anode x-ray tubes". In: *Physics in Medicine & Biology* 54.19 (2009), N433 (cit. on p. 141).

[Prümmer, 2006] Marcus Prümmer, Joachim Hornegger, Marcus Pfister, and Arnd Dörfler. "Multi-Modal 2D-3D Non-Rigid Registration". In: *Medical Imaging 2006: Image Processing*. 2006 (cit. on p. 111).

[Raaymakers, 2017] Bas W Raaymakers, Irene M Jurgenliemk-Schulz, Gerard H Bol, et al. "First patients treated with a 1.5 T MRI-linac: Clinical proof of concept of a high-precision, high-field MRI guided radiotherapy treatment". In: *Physics in Medicine & Biology* 62.23 (2017).

[Radford, 2015] Alec Radford. "Unsupervised representation learning with deep convolutional generative adversarial networks". In: *arXiv preprint arXiv:1511.06434* (2015) (cit. on pp. 57, 59).

[Radiological Protection, 1977] International Commission on Radiological Protection. *Recommendations of the ICRP*. Pergamon Press for the International Commission on Radiological Protection, 1977 (cit. on p. 31).

[Radiology Cafe, 2023] Radiology Cafe. *Production of X-rays - FRCR Physics Notes*. https://www.radiologycafe.com/frcr-physics-notes/x-ray-imaging/production-of-x-rays/. Accessed: 2024-11. 2023 (cit. on p. 41).

[Radon, 2005] Johann Radon. "1.1 über die bestimmung von funktionen durch ihre integralwerte längs gewisser mannigfaltigkeiten". In: *Classic papers in modern diagnostic radiology* 5.21 (2005), p. 124 (cit. on p. 42).

[Rigaud, 2015] B. Rigaud, A. Simon, J. Castelli, and et al. "Evaluation of deformable image registration methods for dose monitoring in head and neck radiotherapy". In: *BioMed Research International* (2015) (cit. on p. 105).

[Rombach, 2022] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. "Latent Diffusion Models for High-Resolution Image Synthesis". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 11784–11795 (cit. on p. 90).

[Ronneberger, 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: *MICCAI*. 2015 (cit. on pp. 31, 82).

[Röntgen, 1895] Wilhelm Conrad Röntgen. "On a new kind of rays". In: *Nature* 53 (1895), pp. 274–276 (cit. on p. 8).

[Ruedinger, 2021] KL Ruedinger, S Schafer, MA Speidel, and CM Strother. "4D-DSA: development and current neurovascular applications". In: *American Journal of Neuroradiology* 42.2 (2021), pp. 214–220 (cit. on p. 165).

[Rusu, 2020] M. Rusu, G. Madoz, J. Faivre, and et al. "Registration of presurgical MRI and histopathology images from radical prostatectomy via RAPSODI". In: *Medical Physics* 47.9 (2020), pp. 4177–4188 (cit. on p. 106).

[Schlemper, 2019] Jo Schlemper, Ozan Oktay, Michiel Schaap, Mattias Heinrich, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. "Attention gated networks: Learning to leverage salient regions in medical images". In: *Medical image analysis* 53 (2019), pp. 197–207 (cit. on p. 113).

[Schwartz, 2013] David L Schwartz, Adam S Garden, Shalin J Shah, et al. "Adaptive radiotherapy for head and neck cancer—Dosimetric results from a prospective clinical trial". In: *Radiotherapy and Oncology* 106.1 (2013), pp. 80–84 (cit. on pp. 22, 24).

[Settecase, 2021] Fabio Settecase and Vitaliy L Rayz. "Advanced vascular imaging techniques". In: *Handbook of Clinical Neurology* 176 (2021), pp. 81–105 (cit. on pp. 165, 166).

[Shang, 2024] Shuyao Shang, Zhengyang Shan, Guangxing Liu, LunQian Wang, XingHua Wang, Zekai Zhang, and Jinglin Zhang. "Resdiff: Combining cnn and diffusion model for image super-resolution". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. 8. 2024, pp. 8975–8983 (cit. on p. 71).

[Shannon, 1949] Claude E. Shannon. "Communication in the Presence of Noise". In: *Proceedings of the IRE* 37.1 (1949), pp. 10–21 (cit. on p. 51).

[Sharp, 2014] Greg Sharp, Karl-Dieter Fritscher, Vladimir Pekar, et al. "Vision 20/20: Perspectives on automated image segmentation for radiotherapy". In: *Medical Physics* 41.5 (2014), p. 050902 (cit. on p. 28).

[Shen, 2022a] Liyue Shen, John Pauly, and Lei Xing. "NeRP: Implicit Neural Representation Learning with Prior EMbedding for Sparsely Sampled Image Reconstruction". In: *IEEE Transactions on Neural Networks* (2022) (cit. on pp. 75–77, 84, 108, 110, 111, 120, 121).

[Shen, 2022b] Liyue Shen, Wei Zhao, Dante Capaldi, John Pauly, and Lei Xing. "A Geometry-Informed Deep Learning Framework for Ultra-Sparse 3D Tomographic Image Reconstruction". In: *Computers in Biology and Medicine* (2022) (cit. on pp. 73, 75, 84, 108, 110, 111).

[Shen, 2019] Liyue Shen, Wei Zhao, and Lei Xing. "Patient-Specific Reconstruction of Volumetric Computed Tomography Images from a Single Projection View via Deep Learning". In: *Nature* 3.11 (2019) (cit. on pp. 73, 74, 108, 110, 111).

[Shepp, 1974] Lawrence A Shepp and Benjamin F Logan. "The Fourier reconstruction of a head section". In: *IEEE Transactions on nuclear science* 21.3 (1974), pp. 21–43 (cit. on p. 44).

[Shepp, 1982] Lawrence A Shepp and Yehuda Vardi. "Maximum likelihood reconstruction for emission tomography". In: *IEEE transactions on medical imaging* 1.2 (1982), pp. 113–122 (cit. on p. 48).

[Shibata, 2021] Hisaichi Shibata, Shouhei Hanaoka, Yukihiro Nomura, Takahiro Nakao, Tomomi Takenaga, Naoto Hayashi, and Osamu Abe. "X2CT-FLOW: Maximum a Posteriori Reconstruction Using a Progressive Flow-Based Deep Generative Model for Ultra Sparse-View Computed Tomography in Ultra Low-Dose Protocols". In: *arXiv*. 2021.

[Shibata, 2022] Hisaichi Shibata, Shouhei Hanaoka, Yukihiro Nomura, Takahiro Nakao, Tomomi Takenaga, Naoto Hayashi, and Osamu Abe. "On the Simulation of Ultra-Sparse-View and Ultra-Low-Dose Computed Tomography with Maximum a Posteriori Reconstruction Using a Progressive Flow-Based Deep Generative Model". In: *Tomography* (2022).

[Siddon, 1985] Robert L. Siddon. "Fast calculation of the exact radiological path for a three-dimensional CT array". In: *Medical Physics* 12.2 (1985), pp. 252–255 (cit. on p. 81).

[Sohl-Dickstein, 2015] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. "Deep unsupervised learning using nonequilibrium thermodynamics". In: *International conference on machine learning*. PMLR. 2015, pp. 2256–2265 (cit. on p. 57).

[Sonke, 2019] Jan-Jakob Sonke, Marianne Aznar, and Coen Rasch. "Adaptive radiotherapy for anatomical changes". In: *Seminars in Radiation Oncology* 29.3 (2019), pp. 245–257 (cit. on pp. 20, 22–25, 27, 30, 31).

[Sonke, 2009] Jan-Jakob Sonke, Lambert Zijp, Peter Remeijer, and Marcel Herk. "Cone-beam computed tomography for radiotherapy positioning and verification". In: *International Journal of Radiation Oncology* Biology* Physics* 73.4 (2009), pp. 1324–1329.

[Spezi, 2012] Emiliano Spezi, Patrick Downes, Richard Jarvis, Emil Radu, and John Staffurth. "Patient-specific three-dimensional concomitant dose from cone beam computed tomography exposure in image-guided radiotherapy". In: *International Journal of Radiation Oncology* Biology* Physics* 83.1 (2012), pp. 419–426 (cit. on p. 17).

[Spinat, 2024] Quentin Spinat, Despoina Ioannidou, Kumar Shreshtha, Ayoub Oumani, Alexandre Cafaro, Olivier Teboul, and Nikos Paragios. *Method for generating a 3d image of a human body part*. US Patent App. 18/635,312. Oct. 2024.

[Sun, 2024a] Jiawei Sun, Nannan Cao, Hui Bi, Liugang Gao, Kai Xie, Tao Lin, Jianfeng Sui, and Xinye Ni. "DiffRecon: Diffusion-based CT reconstruction with cross-modal deformable fusion for DR-guided non-coplanar radiotherapy". In: *Computers in Biology and Medicine* 179 (2024), p. 108868 (cit. on p. 126).

[Sun, 2022] Li Sun, Junxiang Chen, Yanwu Xu, Mingming Gong, Ke Yu, and Kayhan Batmanghelich. "Hierarchical Amortized Gan for 3D High Resolution Medical Image Synthesis". In: *IEEE journal of biomedical and health informatics* 26.8 (2022) (cit. on p. 90).

[Sun, 2024b] Yiran Sun, Hana Baroudi, Tucker Netherton, Laurence Court, Osama Mawlawi, Ashok Veeraraghavan, and Guha Balakrishnan. "DIFR3CT: Latent Diffusion for Probabilistic 3D CT Reconstruction from Few Planar X-Rays". In: *arXiv preprint arXiv:2408.15118* (2024) (cit. on p. 89).

[Tan, 2022] Zhiqiang Tan, Jun Li, Huiren Tao, Shibo Li, and Ying Hu. "XctNet: Reconstruction Network of Volumetric Images from a Single X-Ray Image". In: *Computerized Medical Imaging and Graphics* 98 (2022) (cit. on pp. 73, 75, 108, 110, 111).

[Tan, 2023] Zhiqiang Tan, Shibo Li, Ying Hu, Huiren Tao, and Lihai Zhang. "Semi-XctNet: Volumetric Images Reconstruction Network from a Single Projection Image via Semi-Supervised Learning". In: *Computers in Biology and Medicine* 155 (2023) (cit. on pp. 75, 108, 110, 111).

[Tanner, 2018] Clemens Tanner, Faruk Ozdemir, Robert Profanter, et al. "Generative adversarial networks for MR-CT deformable image registration". In: *arXiv preprint arXiv:1807.07349* (2018) (cit. on p. 104).

[Teguh, 2011] David N Teguh, Peter C Levendag, Pieter W Voet, et al. "Clinical validation of atlas-based auto-segmentation of multiple target volumes and normal tissue (swallowing/mastication) structures in the head and neck". In: *International Journal of Radiation Oncology* Biology* Physics* 81.4 (2011), pp. 950–957 (cit. on p. 28).

[Thibault, 2007] Jean-Baptiste Thibault, Ken D Sauer, Charles A Bouman, and Jiang Hsieh. "A three-dimensional statistical approach to improved image quality for multislice helical CT". In: *Medical physics* 34.11 (2007), pp. 4526–4544 (cit. on p. 49).

[Thing, 2016] Rune Soelberg Thing, Uffe Bernchou, Ernesto Mainegra-Hing, et al. "Hounsfield unit recovery in clinical cone beam CT images of the thorax acquired for image guided radiation therapy". In: *Physics in Medicine & Biology* 61.15 (2016), pp. 5781–5802 (cit. on pp. 17, 29).

[Thirion, 1998] J-P Thirion. "Image matching as a diffusion process: an analogy with Maxwell's demons". In: *Medical image analysis* 2.3 (1998), pp. 243–260 (cit. on p. 105).

[Thörnqvist, 2016] Sara Thörnqvist, Liv B Hysing, Laura Tuomikoski, et al. "Adaptive radiotherapy strategies for pelvic tumors—a systematic review of clinical implementations". In: *Acta Oncologica* 55.8 (2016), pp. 943–958.

[Thummerer, 2020] Adrian Thummerer, Paolo Zaffino, Arturs Meijers, Gabriel Guterres Marmitt, Joao Seco, Roel JHM Steenbakkers, Johannes A Langendijk, Stefan Both, Maria F Spadea, and Antje C Knopf. "Comparison of CBCT based synthetic CT methods suitable for proton dose calculations in adaptive proton therapy". In: *Physics in Medicine & Biology* 65.9 (2020), p. 095002 (cit. on p. 31).

[Tian, 2022] Lin Tian, Yueh Z. Lee, Raúl San José Estépar, and Marc Niethammer. "LiftReg: Limited Angle 2D/3D Deformable Registration". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2022 (cit. on pp. 108, 111–113).

[Tian, 2020] Lin Tian, Connor Puett, Peirong Liu, Zhengyang Shen, Stephen R. Aylward, Yueh Z. Lee, and Marc Niethammer. "Fluid Registration Between Lung CT and Stationary Chest Tomosynthesis Images". In: *MICCAI*. 2020 (cit. on p. 111).

[Tian, 2011] Zhen Tian, Xun Jia, Kaijun Yuan, et al. "Low-dose CT reconstruction via edge-preserving total variation regularization". In: *Physics in Medicine & Biology* 56.18 (2011), pp. 5949–5967 (cit. on p. 31).

[Tibshirani, 1996] Robert Tibshirani. "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 58.1 (1996), pp. 267–288 (cit. on p. 52).

[Tomazevic, 2003] D. Tomazevic, B. Likar, T. Slivnik, and F. Pernus. "3-D/2-D registration of CT and MR to X-ray images". In: *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings*. IEEE. 2003, pp. II–II (cit. on p. 107).

[Ulyanov, 2018] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. "Deep image prior". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 9446–9454 (cit. on pp. 60, 76).

[Unberath, 2018] Mathias Unberath, Jan-Nico Zaech, Sing Chun Lee, Bastian Bier, Javad Fotouhi, Mehran Armand, and Nassir Navab. "DeepDRR–A Catalyst for Machine Learning in Fluoroscopy-Guided Procedures". In: *MICCAI*. 2018 (cit. on pp. 40, 81, 145, 158).

[Vallières, 2020] Martin Vallières, Emily Kay-Rivest abd Léo Jean Perrin, Xavier Liem, Christophe Furstoss, Nader Khaouam, Phuc Félix Nguyen-Tan, Chang-Shu Wang, and Khalil Sultanem. "Data from Head-Neck-PET-CT". In: (2020) (cit. on p. 82).

[Vamvakeros, 2017] Antonios Vamvakeros. "Operando chemical tomography of packed bed and membrane reactors for methane processing". PhD thesis. UCL (University College London), 2017 (cit. on p. 43).

[Van Den Oord, 2017] Aaron Van Den Oord, Oriol Vinyals, et al. "Neural discrete representation learning". In: *Advances in neural information processing systems* 30 (2017) (cit. on p. 90).

[Vaswani, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention Is All You Need". In: *NeurIPS*. 2017 (cit. on p. 115).

[Veiga, 2015] C. Veiga, A. M. Lourenço, S. Mouinuddin, and et al. "Toward adaptive radiotherapy for head and neck patients: uncertainties in dose warping due to the choice of deformable registration algorithm". In: *Medical Physics* (2015) (cit. on p. 105).

[ViewRay Incorporated, 2024] ViewRay Incorporated. *ViewRay: MRI-Guided Radiation Therapy Systems*. https://viewraysystems.com/. Accessed: 2024. 2024 (cit. on p. 19).

[Vision RT, 2024] Vision RT. *AlignRT: Solution de Positionnement et Suivi sans Contact pour la Radiothérapie*. https://visionrt.com/fr/solutions-alignrt/. Accessed: 2024-11-07. 2024 (cit. on p. 159).

[Wang, 2018a] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. "Esrgan: Enhanced super-resolution generative adversarial networks". In: *Proceedings of the European conference on computer vision (ECCV) workshops*. 2018, pp. 0–0 (cit. on p. 70).

[Wang, 2018b] Ying-Cong Wang, Ying Tai, Xiaoming Shao, Jilin Liu, Chengjie Li, and Feiyue Huang. "FSRGAN: Learning face super-resolution guided by 3D facial priors". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 725–741 (cit. on p. 65).

[Wang, 2023] Yufeng Wang and Qing Xia. "TPG-rayGAN: CT Reconstruction Based on Transformer and Generative Adversarial Networks". In: *Third International Conference on Intelligent Computing and Human-Computer Interaction (ICHCI 2022)*. 2023 (cit. on pp. 73, 75, 108, 110, 111).

[Wang, 2003] Zhou Wang, Eero P. Simoncelli, and Alan C. Bovik. "Multiscale Structural Similarity for Image Quality Assessment". In: *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers*. 2003 (cit. on p. 83).

[Weizman, 2016] L. Weizman, Y. C. Eldar, and D. Ben Bashat. "Reference-based MRI". In: *Medical Physics* 43.10 (2016), pp. 5357–5369 (cit. on p. 101).

[Wells III, 1996] William M Wells III, Paul Viola, Hideki Atsumi, Shin Nakajima, and Ron Kikinis. "Multi-modal volume registration by maximization of mutual information". In: *Medical image analysis* 1.1 (1996), pp. 35–51 (cit. on p. 104).

[Weygand, 2016] Joseph Weygand, Clifton David Fuller, Geoffrey S Ibbott, Abdallah SR Mohamed, Yao Ding, Jinzhong Yang, Ken-Pin Hwang, and Jihong Wang. "Spatial precision in magnetic resonance imaging–guided radiation therapy: the role of geometric distortion". In: *International Journal of Radiation Oncology* Biology* Physics* 95.4 (2016), pp. 1304–1316 (cit. on p. 30).

[Wu, 2023] Sean Wu, Naoki Kaneko, Steve Mendoza, David S Liebeskind, and Fabien Scalzo. "3D Reconstruction from 2D Cerebral Angiograms

as a Volumetric Denoising Problem". In: *International Symposium on Visual Computing*. Springer. 2023, pp. 382–393 (cit. on pp. 166, 172).

[Wu, 2015] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. "3d shapenets: A deep representation for volumetric shapes". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 1912–1920 (cit. on p. 70).

[Xia, 2022] Weihao Xia, Yulun Zhang, Yujiu Yang, Jing-Hao Xue, Bolei Zhou, and Ming-Hsuan Yang. "Gan inversion: A survey". In: *IEEE transactions on pattern analysis and machine intelligence* 45.3 (2022), pp. 3121–3138.

[Xu, 2015] Yuan Xu, Ti Bai, Hao Yan, Luo Ouyang, Arnold Pompos, Jing Wang, Linghong Zhou, Steve B Jiang, and Xun Jia. "A practical cone-beam CT scatter correction method with optimized Monte Carlo simulations for image-guided radiation therapy". In: *Physics in Medicine & Biology* 60.9 (2015), p. 3567.

[Yan, 2010] Di Yan. "Adaptive radiotherapy: Merging principle into clinical practice". In: *Seminars in Radiation Oncology* 20.2 (2010), pp. 79–83.

[Yan, 1997] Di Yan, David Lockman, Alvaro Martinez, John Wong, and J. Liang. "A New Concept for Dose-Guided Radiation Therapy: Adaptive Radiation Therapy (ART)". In: *Medical Physics* 24.6 (1997), pp. 907–915 (cit. on p. 20).

[Yan, 2003] Hui Yan, Fang-Fang Yin, and Jae Ho Kim. "A phantom study on the positioning accuracy of the Novalis Body system". In: *Medical physics* 30.12 (2003), pp. 3052–3060.

[Yin, 2002] Fang-Fang Yin, Samuel Ryu, Munther Ajlouni, Jingeng Zhu, Hui Yan, Harrison Guan, Kathleen Faber, Jack Rock, Muwaffak Abdalhak, Lisa Rogers, et al. "A technique of intensity-modulated radiosurgery (IMRS) for spinal tumors". In: *Medical physics* 29.12 (2002), pp. 2815–2822.

[Yin, 2013] Wen-Jing Yin, Ying Sun, Feng Chi, Jian-Lan Fang, Rui Guo, Xiao-Li Yu, Yan-Ping Mao, Zhen-Yu Qi, Ying Guo, Meng-Zhong Liu, et al. "Evaluation of inter-fraction and intra-fraction errors during volumetric modulated arc therapy in nasopharyngeal carcinoma patients". In: *Radiation Oncology* 8 (2013), pp. 1–8 (cit. on p. 16).

[Ying, 2019] Xingde Ying, Heng Guo, Kai Ma, Jian Wu, Zhengxin Weng, and Yefeng Zheng. "X2CT-GAN: Reconstructing CT from Biplanar X-Rays with Generative Adversarial Networks". In: *CVPR*. 2019 (cit. on pp. 73–75, 77, 84, 108, 110, 111, 145).

[Yu, 2021] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. "Pixelnerf: Neural Radiance Fields from One or Few Images". In: *CVPR*. 2021.

[Yu, 2018] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. "Generative image inpainting with contextual attention". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 5505–5514 (cit. on p. 54).

[Yu, 2017] Weimin Yu, Moritz Tannast, and Guoyan Zheng. "Non-Rigid Free-Form 2D–3D Registration Using a B-Spline-Based Statistical Deformation Model". In: *Pattern recognition* 63 (2017).

[Yuan, 2020] Nimu Yuan, Brandon Dyer, Shyam Rao, Quan Chen, Stanley Benedict, Lu Shang, Yan Kang, Jinyi Qi, and Yi Rong. "Convolutional neural network enhancement of fast-scan low-dose cone-beam CT images for head and neck radiotherapy". In: *Physics in Medicine & Biology* 65.3 (2020), p. 035003 (cit. on p. 31).

[Zha, 2022] Ruyi Zha, Yanhao Zhang, and Hongdong Li. "NAF: Neural Attenuation Fields for Sparse-View Cbct Reconstruction". In: *MICCAI*. 2022 (cit. on pp. 75, 76, 111).

[Zhang, 2023a] Baochang Zhang, Shahrooz Faghihroohi, Mohammad Farid Azampour, Shuting Liu, Reza Ghotbi, Heribert Schunkert, and Nassir Navab. "A Patient-Specific Self-Supervised Model for Automatic X-Ray/CT Registration". In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2023*. Springer, 2023 (cit. on p. 107).

[Zhang, 2023b] Chulong Zhang, Jingjing Dai, Tangsheng Wang, Xuan Liu, Yinping Chan, Lin Liu, Wenfeng He, Yaoqin Xie, and Xiaokun Liang. "XTransCT: Ultra-Fast Volumetric CT Reconstruction Using Two Orthogonal X-Ray Projections via a Transformer Network". In: *arXiv*. 2023 (cit. on pp. 73, 75, 108, 110, 111).

[Zhang, 2017] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. "Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising". In: *IEEE transactions on image processing* 26.7 (2017), pp. 3142–3155 (cit. on p. 54).

[Zhang, 2018] L. Zhang, Z. Wang, C. Shi, and et al. "The impact of robustness of deformable image registration on contour propagation and dose accumulation for head and neck adaptive radiotherapy". In: *Journal of Applied Clinical Medical Physics* (2018) (cit. on p. 105).

[Zhang, 2023c] Lvmin Zhang and Maneesh Agrawala. "Adding Conditional Control to Text-to-Image Diffusion Models". In: *arXiv preprint arXiv:2302.05543* (2023) (cit. on p. 128).

[Zhang, 2021a] You Zhang. "An Unsupervised 2D–3D Deformable Registration Network (2D3D-RegNet) for Cone-Beam CT Estimation". In: *Physics in Medicine & Biology* 66.7 (2021) (cit. on pp. 108, 111, 112).

[Zhang, 2019] You Zhang, Michael R. Folkert, Bin Li, Xiaokun Huang, Jeffrey J. Meyer, Tsuicheng Chiu, Pam Lee, Joubin Nasehi Tehrani, Jing Cai, David Parsons, et al. "4D Liver Tumor Localization Using Cone-Beam Projections and a Biomechanical Model". In: *Radiotherapy and Oncology* 133 (2019).

[Zhang, 2016] You Zhang, Joubin Nasehi Tehrani, and Jing Wang. "A Biomechanical Modeling Guided CBCT Estimation Technique". In: 36.2 (2016).

[Zhang, 2021b] Yungeng Zhang, Haifang Qin, Peixin Li, Yuru Pei, Yuke Guo, Tianmin Xu, and Hongbin Zha. "Deformable Registration of Lateral Cephalogram and Cone-Beam Computed Tomography Image". In: *Medical Physics* 48.11 (2021).

[Zhao, 2022] Huangxuan Zhao, Zhenghong Zhou, Feihong Wu, Dongqiao Xiang, Hui Zhao, Wei Zhang, Lin Li, Zhong Li, Jia Huang, Hongyao Hu, et al. "Self-supervised learning enables 3D digital subtraction angiography reconstruction from ultra-sparse 2D projection views: a multicenter study". In: *Cell Reports Medicine* 3.10 (2022) (cit. on p. 166).

[Zheng, 2010] Guoyan Zheng. "Effective Incorporating Spatial Information in a Mutual Information Based 3D–2D Registration of a CT Volume to X-Ray Images". In: *Computerized medical imaging and graphics* 34.7 (2010).

[Zheng, 2021] Yuanjie Zheng, Xiaodan Sui, Yanyun Jiang, Tongtong Che, Shaoting Zhang, Jie Yang, and Hongsheng Li. "SymReg-GAN: symmetric image registration with generative adversarial networks". In: *IEEE transactions on pattern analysis and machine intelligence* 44.9 (2021), pp. 5631–5646 (cit. on p. 104).

[Zhou, 2023] Zhenghong Zhou, Huangxuan Zhao, Jiemin Fang, Dongqiao Xiang, Lei Chen, Lingxia Wu, Feihong Wu, Wenyu Liu, Chuansheng Zheng, and Xinggang Wang. "TiAVox: Time-aware Attenuation Voxels for

Sparse-view 4D DSA Reconstruction". In: *arXiv preprint arXiv:2309.02318* (2023).

[Zhu, 2017] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. "Unpaired image-to-image translation using cycle-consistent adversarial networks". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2223–2232 (cit. on pp. 31, 104).

[Zhu, 2019] Weike Zhu, Yuankai Huang, Li Zeng, et al. "AnatomyNet: Deep learning for fast and fully automated whole-volume segmentation of head and neck anatomy". In: *Medical Physics* 46.2 (2019), pp. 576–589 (cit. on p. 28).

[Zikic, 2008] Darko Zikic, Martin Groher, Ali Khamene, and Nassir Navab. "Deformable Registration of 3D Vessel Structures to a Single Projection Image". In: *Medical imaging 2008: image processing*. 2008 (cit. on p. 111).

[Zollei, 2001a] L Zollei, Eric Grimson, Alexander Norbash, and W Wells. "2D-3D rigid registration of X-ray fluoroscopy and CT images using mutual information and sparsely sampled histogram estimators". In: *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*. Vol. 2. IEEE. 2001, pp. II–II.

[Zollei, 2001b] L. Zollei, E. Grimson, A. Norbash, and W. Wells. "2D-3D rigid registration of X-ray fluoroscopy and CT images using mutual information and sparsely sampled histogram estimators". In: *2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. 2001, pp. 696–703 (cit. on p. 107).

[Zöllner, 2017] Christoph Zöllner, Simon Rit, Christopher Kurz, Gloria Vilches-Freixas, Florian Kamp, George Dedes, Claus Belka, Katia Parodi, and Guillaume Landry. "Decomposing a prior-CT-based cone-beam CT projection correction algorithm into scatter and beam hardening components". In: *Physics and Imaging in Radiation Oncology* 3 (2017), pp. 49–52 (cit. on p. 145).

[Zuley, 2015] M. L. Zuley, R. Jarosz, S. Kirk, Y. Lee, R. Colen, et al. "The Cancer Genome Atlas Head-Neck Squamous Cell Carcinoma Collection TCGA-HNSC)". In: (2015) (cit. on p. 82).

[Zuo, 2021] Jingyi Zuo. "2D to 3D Neurovascular Reconstruction from Biplane View via Deep Learning". In: *2021 2nd International Conference on Computing and Data Science (CDS)*. IEEE. 2021, pp. 383–387 (cit. on p. 166).

université PARIS-SACLAY

ÉCOLE DOCTORALE
Cancérologie : biologie - médecine - santé (CBMS)

Titre: Radiothérapie adaptative guidée par l'intelligence artificielle pour les cancers ORL

Mots clés: Intelligence Artificielle - Vision par Ordinateur - Analyse d'Images Médicales - Radiothérapie adaptative Cancer de la Tête et du Cou - Problème inverse

Résumé: Le cancer de la tête et du cou (HNC) est l'un des cancers les plus difficiles à traiter en raison de la complexité de son anatomie et des changements significatifs spécifiques à chaque patient au cours du traitement. En tant que 6e cancer le plus fréquent dans le monde, le HNC présente souvent un mauvais pronostic en raison d'un diagnostic tardif et de l'absence de marqueurs prédictifs fiables. La radiothérapie, souvent associée à la chirurgie, est confrontée à des défis tels que la variabilité inter-observateur, la complexité de la planification et les changements anatomiques pendant le traitement.

La radiothérapie adaptative est essentielle pour maintenir la précision à mesure que l'anatomie du patient évolue. Cependant, les méthodes d'imagerie peu invasives actuelles, comme la tomographie conique (CBCT) et les rayons X biplanaires, sont limitées en qualité ou ne fournissent que des images 2D, ce qui complique l'adaptation quotidienne du traitement. Cette thèse propose des approches innovantes basées sur l'apprentissage profond pour reconstruire des images CT 3D précises à partir de rayons X biplanaires, permettant une radiothérapie adaptative qui réduit la dose de radiation, accélère l'acquisition, réduit les coûts et améliore la précision.

La reconstruction de volumes 3D à partir de rayons X biplanaires est difficile en raison des informations limitées de seulement deux projections, ce qui crée une ambiguïté importante dans la capture des structures internes. Pour y remédier, cette thèse intègre des a priori anatomiques et de déformation via l'apprentissage profond, améliorant ainsi considérablement la précision des reconstructions malgré des données limitées.

La première méthode, X2Vision, est une approche non supervisée qui utilise des modèles génératifs entraînés sur des scans CT pour apprendre la distribution des anatomies de la tête et du cou. Elle optimise des vecteurs latents pour générer des volumes 3D alignés avec les rayons X biplanaires et les a priori anatomiques. En utilisant ces a priori et en naviguant dans le domaine anatomique, X2Vision réduit considérablement la nature mal posée du problème de reconstruction, obtenant des résultats précis même avec seulement deux projections.

En radiothérapie, des scans pré-traitement comme le CT ou l'IRM sont souvent disponibles et essentiels pour améliorer les reconstructions en tenant compte des changements anatomiques au fil du temps. Nous avons développé XSynthMorph, une méthode qui intègre des caractéristiques spécifiques au patient à partir des scans CT préalablement acquis. En combinant des a priori anatomiques et de déformation, XSynthMorph s'adapte aux changements tels que la perte de poids ou les déformations non rigides, permettant des reconstructions plus robustes et personnalisées, avec une précision et un détail sans précédent.

Nous avons exploré le potentiel clinique de X2Vision et XSynthMorph, avec des évaluations cliniques préliminaires montrant leur efficacité dans le positionnement du patient, la recnstruction des structures et l'analyse dosimétrique, soulignant leur potentiel pour la radiothérapie adaptative quotidienne. Pour approcher la réalité clinique, nous avons développé une première approche pour intégrer ces méthodes aux systèmes de rayons X biplanaires utilisés en radiothérapie.

En conclusion, cette thèse démontre la faisabilité de la radiothérapie adaptative utilisant uniquement des rayons X biplanaires. En combinant des modèles génératifs, des a priori de déformation et des scans préalablement acquis, nous avons montré que des reconstructions 3D de haute qualité peuvent être obtenues avec une faible exposition aux radiations. Ce travail ouvre la voie à une radiothérapie adaptative quotidienne, offrant une solution peu invasive, peu coûteuse, et précise.

université PARIS-SACLAY | ÉCOLE DOCTORALE Cancérologie : biologie - médecine - santé (CBMS)

Title: AI-Driven Adaptive Radiation Treatment Delivery for Head & Neck Cancers

Keywords: Artificial Intelligence - Computer Vision - Medical Image Analysis - Adaptative Radiotherapy - Head and Neck Cancer - Inverse Problem

Abstract: Head and neck cancer (HNC) is one of the most challenging cancers to treat due to its complex anatomy and significant patient-specific changes during treatment. As the 6^{th} most common cancer worldwide, HNC often has a poor prognosis due to late diagnosis and the lack of reliable predictive markers. Radiation therapy, typically combined with surgery, faces challenges such as inter-observer variability, complex treatment planning, and anatomical changes throughout the treatment process.

Adaptive radiotherapy is essential to maintain precision as the patient's anatomy evolves during treatment. However, current low-invasive imaging methods before each treatment fraction, such as Cone Beam CT (CBCT) and biplanar X-rays, are limited in quality or provide only 2D images, making daily treatment adaptation challenging. This thesis introduces novel deep learning approaches to reconstruct accurate 3D CT images from biplanar X-rays, enabling adaptive radiotherapy that reduces radiation dose, shortens acquisition times, lowers costs, and improves treatment precision.

Reconstructing 3D volumes from biplanar X-rays is inherently challenging due to the limited information provided by only two projections, leading to significant ambiguity in capturing internal structures. To address this, the thesis incorporates anatomical and deformation priors through deep learning, significantly improving reconstruction accuracy despite the very sparse measurements.

The first method, X2Vision, is an unsupervised approach that uses generative models trained on head and neck CT scans to learn the distribution of head and neck anatomies. It optimizes latent vectors to generate 3D volumes that align with both biplanar X-rays and anatomical priors. By leveraging these priors and navigating the anatomical manifold, X2Vision dramatically reduces the ill-posed nature of the reconstruction problem, achieving accurate results even with just two projections.

In radiotherapy, pre-treatment scans such as CT or MRI are typically available and are essential for improving reconstructions by accounting for anatomical changes over time. To make use of this data, we developed XSynthMorph, a method that integrates patient-specific features from pre-acquired planning CT scans. By combining anatomical and deformation priors, XSynthMorph adjusts for changes like weight loss, non-rigid deformations, or tumor regression. This approach enables more robust and personalized reconstructions, providing an unprecedented level of precision and detail in capturing 3D structures.

We explored the clinical potential of X2Vision and XSynthMorph, with preliminary clinical evaluations demonstrating their effectiveness in patient positioning, structure retrieval, and dosimetry analysis, highlighting their promise for daily adaptive radiotherapy. To bring these methods closer to clinical reality, we developed an initial approach to integrate them into real-world biplanar X-ray systems used in radiotherapy.

In conclusion, this thesis demonstrates the feasibility of adaptive radiotherapy using only biplanar X-rays. By combining generative models, deformation priors, and pre-acquired scans, we have shown that high-quality 3D reconstructions can be achieved with minimal radiation exposure. This work paves the way for daily adaptive radiotherapy, offering a low-invasive, cost-effective solution that enhances precision, reduces radiation exposure, and improves overall treatment efficiency.