*A project report on*

**LungCraft: 3D Modeling and Visualization for Enhanced Diagnosis of Lung Cancer Using CT Scans**

*Submitted in partial fulfillment for the award of the degree of*

**M.Tech. (Integrated) Computer Science and Engineering with Specialization in Business Analytics**

*by*

**AKILA B (20MIA1052)**



**SCHOOL OF COMPUTER SCIENCE AND ENGINEERING**

April, 2024



**DECLARATION**

I hereby declare that the thesis entitled “SINGLE-VIEW DEPTH ESTIMATION: ADVANCING 3D SCENE INTERPRETATION WITH ONE LENS ” submitted by me ,for the award of the degree of M.Tech. (Integrated) Computer Science and Engineering with Specialization in Business Analytics, Vellore Institute of Technology, Chennai, is are cord of Bonafide work carried out by me under the supervision of Dr. Kavitha

I further declare that the work reported in this thesis has not been submitted and will not be submitted, either in part or in full, for the award of any other degree or diploma in this institute or any other institute or university.

**Place: Chennai**

**Date:** **Signature of the Candidate**



**School of Computer Science and Engineering**

CERTIFICATE

This is to certify that the report entitled **“SINGLE-VIEW DEPTH ESTIMATION: ADVANCING 3D SCENE INTERPRETATION WITH ONE LENS”** is prepared and submitted by **Akila B (20MIA1052)** to Vellore Institute of Technology, Chennai, in partial fulfillment of the requirement for the award of the degree of **M.Tech. (Integrated) Computer Science and Engineering with Specialization in Business Analytics** programme is a Bonafide record carried out under my guidance. The project fulfills the requirements as per the regulations of this University and in my opinion meets the necessary standards for submission. The contents of this report have not been submitted and will not be submitted either in part or in full, for the award of any other degree or diploma and the same is certified.

Signature of the Guide:

Name: Dr. D. Kavitha

Date:

Signature of the Examiner 1 Signature of the Examiner 2

Name: Dr. Suganeshwari Name: Dr. Sudha

Date: Date:

Approved by the Head of Department

**ABSTRACT**

LungCraft is a new application of the advanced 3D modeling and visualization techniques for computed tomography (CT) scans to better lung cancer diagnosis. In this study, a detailed dataset from The Cancer Imaging Archive will be applied by diagnostic contrast-enhanced CT scans from 61 patients diagnosed with lung adenocarcinoma. In general, the objective is to evaluate prognosis more accurately by quantitatively analyzing how image features relate to the characteristics of tumors and their outcomes in patients. The two features implemented in this work were most critical to CT: intratumor heterogeneity and tumor shape complexity, systematically scored from routinely acquired diagnostic CT images. Such features allow distinct imaging phenotypes to be recognized and are associated with highly correlated survival differences. LungCraft applies static and interactive 3D visualizations to achieve a more intuitive presentation of the morphology of tumors in the better perception of tumor behavior than the traditional techniques applied with 2D imaging. The implementation is based on a hybrid modeling approach that balances machine learning algorithms designed for the analysis of CT scans with compensation for variation in clinical image acquisition. Our results show that the quantitative imaging features are reproducible and stable, thereby revalidating their potential as valuable tools for making diagnostic decisions on lung adenocarcinoma management. This work underscores the utility of 3D imaging as an adjunct to conventional diagnostic practices and possibly transforming the approach toward evaluation and management of lung cancer in the future. Future work will include refining these techniques and clinical evaluation to determine how they can be integrated into clinical workflows.



**ACKNOWLEDGEMENT**

It is my pleasure to express with deep sense of gratitude to Dr. D. Kavitha <Designation>, School of Computer Science and Engineering, Vellore Institute of Technology, Chennai, for her constant guidance, continual encouragement, understanding; more than all, she taught me patience in my endeavor. My association with her is not confined to academics only, but it is a great opportunity on my part of work with an intellectual and expert in the field of <area>.

It is with gratitude that I would like to extend my thanks to the visionary leader Dr. G. Viswanathan our Honorable Chancellor, Mr. Sankar Viswanathan, Dr. Sekar Viswanathan, Dr. G V Selvam Vice Presidents, Dr. Sandhya Pentareddy, Executive Director, Ms. Kadhambari S. Viswanathan, Assistant Vice-President, Dr. V. S. Kanchana Bhaaskaran Vice-Chancellor, Dr.T. Thyagarajan Pro-Vice Chancellor, VIT Chennai and Dr. P. K. Manoharan, Additional Registrar for providing an exceptional working environment and inspiring all of us during the tenure of the course.

Special mention to Dr. Ganesan R, Dean, Dr. Parvathi R, Associate Dean Academics, Dr. Geetha S, Associate Dean Research, School of Computer Science and Engineering, Vellore Institute of Technology, Chennai for spending their valuable time and efforts in sharing their knowledge and for helping us in every aspect.

In jubilant state, I express ingeniously my whole-hearted thanks to Dr.Sivabalakrishnan. M, Head of the Department, Project Coordinator, Dr. Yogesh C, SCOPE, Vellore Institute of Technology, Chennai, for their valuable support and encouragement to take up and complete the thesis.

My sincere thanks to all the faculties and staff at Vellore Institute of Technology, Chennai, who helped me acquire the requisite knowledge. I would like to thank my parents for their support. It is indeed a pleasure to thank my friends who encouraged me to take up and complete this task.

**Place: Chennai**

**Date:** **Akila B**

**Contents**

[LIST OF FIGURES 8](#_Toc182002796)

[LIST OF TABLES 9](#_Toc182002797)

[Introduction 1](#_Toc182002798)

[1.2 CHALLENGES IN DIAGNOSING LUNG ADENOCARCINOMA 2](#_Toc182002799)

[1.3 LIMITATIONS OF TRADITIONAL 2D IMAGING IN LUNG CANCER DIAGNOSIS 2](#_Toc182002800)

[1.4 THE ROLE OF TUMOR MORPHOLOGY AND DENSITY IN PROGNOSIS 2](#_Toc182002801)

[1.6 AIMS OF LUNGCRAFT IN ADVANCING DIAGNOSTIC PRECISION AND PERSONALIZED TREATMENT 3](#_Toc182002802)

[Related Work 4](#_Toc182002803)

[Dataset 10](#_Toc182002804)

[3.1 DATA OVERVIEW 10](#_Toc182002805)

[3.2 DATASET SUMMARY 11](#_Toc182002806)

[3.3 DATA ACCESS 11](#_Toc182002807)

[3.4 DICOM FORMAT AND HOUNSFIELD UNITS 12](#_Toc182002808)

[3.5 PREPROCESSING PIPELINE 13](#_Toc182002809)

[3.6 ANNOTATION AND LABELING 14](#_Toc182002810)

[Training and Testing Data 15](#_Toc182002811)

[4.1 DETAILED TRAIN-TEST SPLIT 15](#_Toc182002812)

[4.2 DATA FORMAT 18](#_Toc182002813)

[Methodology 23](#_Toc182002814)

[5.1 DATASET PREPARATION 23](#_Toc182002815)

[5.2 PREPROCESSING 24](#_Toc182002816)

[5.3 MODEL DESIGN 25](#_Toc182002817)

[5.4 TRAINING AND HYPERPARAMETER TUNING 27](#_Toc182002818)

[5.5 EVALUATION 29](#_Toc182002819)

[5.6 POST PROCESSING AND VIZUALIZATION 29](#_Toc182002820)

[2D Training 33](#_Toc182002821)

[6.1 TRAINING DATASET 33](#_Toc182002822)

[6.2 DATA PROCESSING 34](#_Toc182002823)

[6.3 MODEL ARCHITECTURE 36](#_Toc182002824)

[6.4 LOSS FUNCTION 37](#_Toc182002825)

[6.5 OPTIMZATION TECHNIQUES 38](#_Toc182002826)

[6.6 EVALUATION METRICS AND BENCHMARKING 39](#_Toc182002827)

[6.7 POST PROCESSING AND VISUALIZATIONS 39](#_Toc182002828)

[3D Training 40](#_Toc182002829)

[7.1 3D DATA AND DATASET 40](#_Toc182002830)

[7.2 MODEL ARCHITECTURE 41](#_Toc182002831)

[7.3 LOSS FUNCTION 42](#_Toc182002832)

[7.4 OPTIMIZATION TECHNIQUES 44](#_Toc182002833)

[7.5 EVALUATION METRICS AND BENCHMARKING 45](#_Toc182002834)

[Results and Discussion 46](#_Toc182002835)

[8.1 SUMMARY TABLE OF THE TRAINING AND VALIDATIONS RESULTS 46](#_Toc182002836)

[8.2 KEY FINDINGS 47](#_Toc182002837)

[8.3 INFERENCES 50](#_Toc182002838)

[Conclusion 51](#_Toc182002839)

[References 54](#_Toc182002840)

**List of Figures**

[Figure 1 LungCT-Diagnosis dataset from Cancer Imaging Archive. 10](#_Toc182002841)

[Figure 2 Visualization of raw DICOM files of different subjects. 11](#_Toc182002842)

[Figure 3 Visualization of a slice of CT scan of the subject R\_210 showing raw metadata attached to it. 12](#_Toc182002843)

[Figure 4 DICOM Slice visualized without DICOM Viewer using pydicom 13](#_Toc182002844)

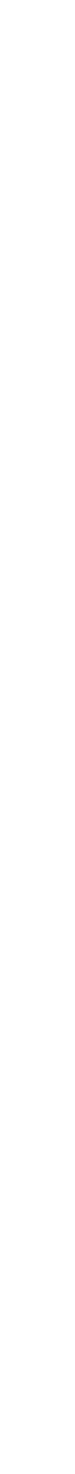
[Figure 5 Data before preprocessing techniques. 13](#_Toc182002845)

### LIST OF TABLES

2.1 MULTIPATH ROUTING PROTOCOLSIN AMNETS 22

* 1. SIMULATIONPARAMETERS 48

**LIST OF ACRONYMS**

MANET Mobile Ad hoc Network WAP Wireless AccessPoint

**Chapter 1**

# Introduction

Lung cancer stands as one of the leading causes of cancer-related deaths worldwide, with lung adenocarcinoma representing the most prevalent and aggressive subtype. Due to its high recurrence rate and complex growth patterns, lung adenocarcinoma presents significant challenges for early diagnosis and prognosis, critical factors for improving patient survival. Despite advancements in imaging technology, current diagnostic approaches—primarily two-dimensional (2D) computed tomography (CT) scans—fall short in capturing the intricate structural and density-based details of these tumors. The limitations inherent in 2D imaging, including a lack of spatial depth and difficulty in detecting intratumor heterogeneity, can lead to misinterpretation of tumor characteristics and suboptimal treatment decisions.

A major obstacle in diagnosing lung adenocarcinoma effectively is its morphological complexity and internal density variability. Traditional imaging techniques struggle to accurately represent these characteristics, which are crucial indicators for prognosis and therapy planning. Tumor shape complexity and density fluctuations, for example, have been associated with patient outcomes and disease progression. However, 2D CT scans provide minimal information on these dimensions, making it difficult for clinicians to evaluate the full extent of tumor aggressiveness and inform personalized treatment approaches.

This paper introduces **LungCraft**, an innovative framework designed to bridge these diagnostic gaps by offering three-dimensional (3D) modeling and interactive visualization capabilities for lung tumors. LungCraft leverages data from The Cancer Imaging Archive (TCIA), specifically the LungCT-Diagnosis collection, which includes CT scans from 61 patients with associated clinical metadata. This dataset enables comprehensive feature extraction, including tumor shape and density variations, providing a solid foundation for prognostic evaluation. By transforming CT scan data into high-resolution 3D models, LungCraft allows clinicians to examine tumors from multiple angles, thereby gaining deeper insights into morphological and density-based complexities that may influence clinical outcomes.

LungCraft aims to complement and extend traditional diagnostic tools by offering a dynamic and intuitive approach to lung cancer imaging, enabling improved accuracy in diagnosis and prognosis. Through interactive 3D visualization combined with quantitative feature scoring, LungCraft serves as a powerful tool for enhancing understanding of tumor behavior, potentially improving personalized treatment strategies for lung cancer patients.

* 1. GLOBAL BURDEN OF LUNG CANCER AND THE NEED FOR IMPROVED DIAGNOSTIC TOOLS

Lung cancer is one of the leading causes of cancer-related deaths worldwide, accounting for more deaths than colon, breast, and prostate cancers combined. Among the different types of lung cancer, adenocarcinoma is particularly lethal due to its high recurrence rates and complex growth patterns. Early detection and precise diagnosis are crucial for improving patient survival; however, these goals are difficult to achieve with the current standard of care. While advances in imaging technology have enhanced lung cancer detection rates, accurate diagnostic methods that offer detailed insights into tumor morphology and density are still lacking. This underscores the need for innovative tools that can provide a more comprehensive understanding of lung tumors, thereby enabling personalized and effective treatment plans.

## CHALLENGES IN DIAGNOSING LUNG ADENOCARCINOMA

Lung adenocarcinoma poses unique diagnostic challenges due to its heterogeneous nature. Tumors often exhibit a high degree of morphological and density variation, which complicates efforts to classify them accurately and assess their malignancy. The variations within tumors are clinically significant, as they correlate with patient outcomes, aggressiveness of the disease, and response to treatment. Without a thorough understanding of these features, clinicians are limited in their ability to accurately predict disease progression and tailor treatment plans accordingly. Addressing these challenges requires diagnostic tools that can capture and analyze the intricate structures within tumors.

## LIMITATIONS OF TRADITIONAL 2D IMAGING IN LUNG CANCER DIAGNOSIS

Most routine diagnostics for lung cancer rely on two-dimensional (2D) computed tomography (CT) scans, which have several limitations. While 2D CT imaging provides a cross-sectional view of the lung, it fails to capture the full three-dimensional (3D) complexity of tumors. This lack of spatial depth can obscure critical features such as shape irregularities and subtle density variations within the tumor. These features are essential for accurate diagnosis and prognosis but are often missed in 2D imaging, leading to potential misinterpretation of the tumor’s true characteristics. As a result, traditional imaging techniques may not provide the detail needed to support accurate prognostic assessments and optimal treatment strategies.

## THE ROLE OF TUMOR MORPHOLOGY AND DENSITY IN PROGNOSIS

Tumor morphology (shape complexity) and intratumor density variations are crucial biomarkers in lung adenocarcinoma. Studies have shown that tumors with irregular shapes or heterogeneous densities are more likely to be aggressive and associated with poorer patient outcomes. Shape complexity may reflect tumor invasiveness, while density variations may indicate areas of differing cellular composition, necrosis, or increased malignancy. These features provide valuable information for staging the disease, forecasting survival rates, and tailoring therapy. However, the inability of 2D imaging to accurately capture these attributes limits clinicians’ capacity to make well-informed prognoses, creating a demand for a more sophisticated approach that can analyze these parameters in three dimensions.

* 1. INTRODUCING LUNGCRAFT: A 3D MODELING AND VISUALIZATION FRAMEWORK

This paper introduces **LungCraft**, a novel framework designed to overcome the limitations of traditional 2D CT imaging. LungCraft provides 3D modeling and interactive visualization capabilities that enable clinicians to observe lung tumors from multiple perspectives. By reconstructing tumor features in 3D, LungCraft enhances clinicians' ability to assess the morphological and density-based complexities of lung tumors with greater accuracy. This advanced framework offers a more intuitive and comprehensive approach to lung cancer diagnostics, supporting early and precise detection while facilitating more informed treatment planning.

## AIMS OF LUNGCRAFT IN ADVANCING DIAGNOSTIC PRECISION AND PERSONALIZED TREATMENT

LungCraft’s primary goal is to enhance diagnostic accuracy and prognostic assessments in lung cancer through advanced 3D imaging and interactive visualization. By creating high-resolution 3D models, LungCraft provides a dynamic and intuitive approach to understanding tumor morphology and internal variations. The framework also integrates quantitative scoring based on tumor shape and density characteristics, which aids in classifying tumors by imaging phenotype and informing clinical decisions. Through this hybrid approach, LungCraft serves as a valuable tool for clinicians, helping them to tailor treatment plans based on a detailed understanding of each patient’s tumor profile, ultimately advancing the goals of personalized medicine.

**Chapter 2**

# Related Work

Gerckens et al. (2019) developed 3D lung tissue cultures (3D- LTCs) from precision-cut lung slices (PCLS) to model human lung diseases more effectively. Their work emphasizes the importance of preserving tissue architecture, biomechanics, and cellular diversity to closely resemble in situ lung conditions. The 3D-LTCs enable detailed analysis of drug responses and disease mechanisms, particularly in fibrosis research, offering personalized insights through patient- derived tissue. While their focus lies on ex vivo modeling for biochemical studies, our research diverges by using non- invasive CT-based 3D models to visualize tumor structures and predict patient outcomes. Both studies contribute to advancing personalized medicine, with Gerckens et al. providing tissue-level insights and our work focusing on diagnostic imaging for improved lung cancer prognosis [1].

Alakwaa et al. (2017) introduced a computer-aided diagnosis (CAD) system leveraging 3D convolutional neural networks (3D-CNNs) for lung cancer detection using CT scans. Their approach focuses on automating the segmentation, detection, and classification processes by integrating a modified U-Net and 3D-CNNs. Although the use of deep learning in their work yields a high classification accuracy of 86.6%, it also highlights the challenges of false positives during nodule detection. While their research centers on the application of machine learning models to automate cancer diagnosis, our work shifts toward enhancing diagnostic visualization through interactive 3D models of CT scans. Both studies aim to improve lung cancer management, but ours focuses on combining imaging phenotypes and survival analysis to offer a more interpretable, non-invasive approach for clinical applications [2].

Cunniff et al. (2021) explored the development and visualization of lung organoids, emphasizing the importance of maintaining the 3D architecture of lung tissue for accurate biological modeling. Their work highlights how conventional 2D culture systems fail to capture the complexity of lung structures, which is essential for understanding disease mechanisms. By leveraging high-resolution microscopy, they demonstrate how lung organoids can replicate in vivo conditions, offering insights into both healthy and diseased states of the lung. While their research focuses on 3D tissue cultures for experimental manipulation, our study extends the idea of 3D modeling by applying it directly to diagnostic CT scans, enabling clinicians to visualize tumor heterogeneity and shape complexity interactively. Both efforts underscore the critical role of 3D visualization in advancing lung disease research and treatment strategies [3].

Uhl et al. (2015) developed and validated patient-derived 3D lung tissue cultures (LTCs) to investigate Wnt/β-catenin signaling for lung tissue repair, focusing on chronic obstructive pulmonary disease (COPD). Their work emphasizes the importance of mimicking the native lung microenvironment through 3D models, which allow for detailed molecular analysis and high-resolution imaging. They demonstrated the activation of Wnt/β-catenin signaling pathways in patient-derived tissues, showing promising therapeutic potential through enhanced epithelial cell regeneration and reduced tissue degradation. While their study focuses on preclinical therapeutic validation for COPD, it aligns with our work’s goal of leveraging 3D visualization for diagnostic and therapeutic advancements. Both approaches highlight the utility of 3D models in advancing personalized healthcare by closely replicating patient- specific lung environments [4].

Tan and Liu (2021) introduced a novel framework that combines 3D CNNs with BERT for the automatic diagnosis of COVID-19 from CT scan images. Their approach involves preprocessing CT slices by segmenting lung regions and filtering out irrelevant backgrounds. A resampling method ensures consistent input size by selecting a fixed number of slice images for training and validation. The framework uses a 3D CNN to extract spatial features, integrated with BERT to enhance contextual understanding of the images. The extracted features are aggregated into a feature vector and fed into a multi-layer perceptron (MLP) for final classification. Their model achieved high performance, with an F1 score of 0.9261 on validation and 0.8822 on test datasets. This work demonstrates the effectiveness of combining deep learning models with language models for image-based medical diagnosis, which aligns with our use of 3D data but emphasizes COVID-19 detection. The framework showcases the potential of 3D models for diagnostic applications, further motivating our exploration of 3D visualization for lung cancer analysis [5].

Ikeda et al. (2013) highlighted the value of three-dimensional (3D) computed tomography (CT) lung modeling in enhancing the precision and safety of lung cancer surgeries. With the rise of minimally invasive procedures, such as video-assisted thoracoscopic surgery (VATS) lobectomy and segmentectomy, understanding the complex anatomy of pulmonary vessels and bronchi is critical for successful operations. The use of multi-detector CT (MDCT) allows surgeons to generate detailed 3D images of lung structures, enabling preoperative simulations and real-time navigation during surgery. This technology provides surgeons with a comprehensive view of vascular and bronchial pathways, enhancing both the accuracy of dissections and patient safety. Advances in imaging software have further improved the visualization of small vessels, supporting more precise segmentectomy procedures. Ikeda et al. emphasize that 3D modeling not only facilitates safer surgeries but also serves as a valuable tool for training junior surgeons and boosting operator confidence. This approach aligns with our focus on leveraging 3D visualization to provide more detailed insights into lung structures, reinforcing the potential for improved medical interventions through advanced modeling techniques [6].

Cheng et al. (2016) explored the transformative impact of three-dimensional (3D) printing and 3D slicer technology in understanding and treating structural lung diseases. As the 3D printing industry advances, clinicians have begun to leverage these technologies for various applications, including pre- procedural planning, biomedical tissue modeling, and the production of custom implantable devices. However, the authors note that despite the growing adoption of rapid prototyping and additive manufacturing techniques in healthcare, many physicians still lack the necessary technical skills to utilize these innovative tools effectively. They discuss the rapid growth of the 3D printing sector, which introduces a multitude of 3D printers and materials, emphasizing the need for clinicians to remain informed about these developments to fully exploit their potential benefits. The paper reviews the history of 3D printing and its recent biomedical applications while addressing the significant barriers to its widespread adoption in the medical field. Cheng et al. also provide a guide for clinicians on designing personalized airway prostheses using 3D Slicer, aiming to facilitate greater participation in the 3D printing sector. This work underscores the importance of integrating advanced technologies into clinical practice, aligning with our objective of enhancing lung cancer diagnosis through innovative modeling and visualization techniques [7].

Dillavou et al. (2003) conducted a study to evaluate the accuracy of aortic diameter measurements using two- dimensional (2D) versus three-dimensional (3D) computed tomography (CT) scans. The research involved two independent, blinded observers who measured the aortic neck and sac diameters from 40 2.5-mm 2D CT scans of 31 patients using electronic calipers and a circular tool for 3D reconstructions. The measurements were obtained at specified anatomical landmarks, with data analyzed through intraclass correlation coefficients (ICC), Bland-Altman variation assessments, and absolute differences. The results indicated a high correlation between 2D minor axis measurements and 3D measurements, with the ICC values for the neck and sac demonstrating strong reliability (neck ICC = 0.9282; sac ICC = 0.8956). While the correlation for the major axis was lower, the average absolute difference between 3D and 2D minor axis diameters was just 1.01 mm, compared to 2.61 mm for the major axis. The authors concluded that minor axis measurements derived from 2D scans are generally sufficient for clinical applications, suggesting that 3D reconstructions may not be necessary for routine aortic diameter assessments. This study highlights the potential for simpler imaging techniques in clinical practice while also setting a foundation for the exploration of more advanced imaging methods, such as 3D modeling, in lung cancer diagnosis [8].

Kumar and Vijai (2012) discuss advancements in 3D reconstruction techniques of the human face from 2D CT scan images, emphasizing their applications in craniofacial surgery. The authors propose a software tool designed to aid surgeons in planning and executing facial reconstruction procedures. They analyze various existing approaches for 3D reconstruction, which range from applications in visualizing real-world scenarios to modeling anatomical structures. The paper evaluates different algorithms, highlighting their suitability for reconstructing various regions of the face, including soft tissues and hard bones. The survey concludes by assessing the effectiveness of each method and recommending the most appropriate approaches for specific medical applications, underscoring the importance of precise modeling in improving surgical outcomes. This work contributes to the growing field of 3D imaging and modeling in medicine, which has significant implications for surgical planning and patient care [9].

Duquette et al. (2012) developed a semi-automatic method for 3D segmentation of the abdominal aorta, focusing on the aneurismal sac of abdominal aortic aneurysms (AAAs) from both CT and MR images. Utilizing graph cut theory, the method minimizes human intervention while effectively segmenting the lumen interface and aortic wall. Tested on a dataset of 44 patients and 10 synthetic images, the segmentation results were compared to manual tracings from four experts, demonstrating that the semi-automatic method achieved similar variability to human operators. This approach provides reliable and reproducible evaluations of the abdominal aorta, enhancing diagnostic accuracy and treatment planning for AAAs [10].

Kabadi et al. (2019) developed a novel 3D lung microtissue model to study nanoparticle-induced alterations in cell- matrix interactions, particularly focusing on multi-walled carbon nanotubes (MWCNTs). Traditional toxicity testing methods rely heavily on 2D in vitro assays and in vivo animal studies, which often yield conflicting results. In contrast, this study co-cultured human lung fibroblasts and epithelial cells with macrophages to form scaffold-free 3D microtissues, which more accurately mimic human physiology. After exposing these microtissues to MWCNTs, carbon black nanoparticles, and crocidolite asbestos fibers, the researchers evaluated microtissue viability, morphology, and gene expression associated with inflammation and extracellular matrix remodeling. The findings indicate that 3D lung microtissues can effectively predict chronic pulmonary endpoints and provide a more relevant alternative for nanomaterial toxicity testing, enhancing understanding of toxicity pathways and potential health hazards from nanoparticle exposure [11].

Subburaj, Ravi, and Agarwal (2009) present a novel computer graphics-based method for the automated identification of anatomical landmarks on 3D bone models reconstructed from CT scan images. The accurate localization of these landmarks is critical for patient-specific preoperative planning, such as tumor referencing and implant alignment, as well as for intra-operative navigation. The authors’ method segments the bone model's surface into different landmark regions—peak, ridge, pit, and ravine—based on surface curvature, and employs an iterative process using a spatial adjacency relationship matrix to label these regions automatically. The performance of the automated system was evaluated against manual landmark identification by three experienced orthopedic surgeons on three 3D bone models, revealing variability in landmark location ranging from 2.15–5.98 mm for the manual method and 1.92–4.88 mm for the automated approach. The results indicate that the automated methodology performed comparably or better than manual identification, demonstrating reproducibility and potential for various applications in surgical planning and navigation [12].

El-Baz et al. (2013) present a novel algorithm for the automatic detection of lung nodules in chest spiral low-dose CT (LDCT) scans, addressing a significant challenge in the computer analysis of chest radiographs. The proposed method involves three main steps: first, it isolates lung nodules, arteries, veins, bronchi, and bronchioles from surrounding anatomical structures. The second step utilizes deformable 2D and 3D templates that characterize the typical geometry and gray-level distribution of lung nodules for detection. This detection process combines normalized cross- correlation template matching with a genetic optimization algorithm to enhance accuracy. In the final step, the algorithm reduces false positives by applying three robust features to distinguish true lung nodules. Testing on 200 CT datasets demonstrates that the proposed algorithm achieves results comparable to those of expert radiologists, highlighting its potential utility in clinical practice for improving lung nodule detection accuracy [13].

Anwar (2021) discusses an innovative approach to diagnosing COVID-19 using AutoML techniques applied to 3D CT scans. The study highlights the limitations of polymerase chain reaction (PCR) tests, which have a high false-negative rate, necessitating alternative diagnostic methods. CT scans provide detailed insights into the chest but typically involve analyzing hundreds of slices, making manual diagnosis by radiologists and pulmonologists time- consuming. To address this, the author proposes an automated AI-based method that leverages AutoML for efficient diagnosis. The model is trained on 2D slices of CT scans rather than 3D scans, with predictions aggregated to label the overall 3D CT scan based on the most frequently occurring diagnosis among the slices. By employing different thresholds, the model classifies scans as COVID-positive or negative. The approach achieved an impressive accuracy of 89% and an F1-score of 88%, demonstrating its potential as a reliable diagnostic tool. The implementation is publicly available, contributing to further research and application in the clinical setting [14].

Serte and Demirel (2021) present a deep learning approach for diagnosing COVID-19 using 3D CT scans, addressing the urgent need for efficient detection methods in hospitals. With the rapid spread of COVID-19 and its associated mortality, timely isolation of infected individuals is crucial. CT scans serve as valuable diagnostic tools, but the extensive number of slices in each scan can cause delays in diagnosis. To overcome this challenge, the authors propose an AI-based methodology utilizing the ResNet-50 deep learning model to classify CT images as either COVID-19 positive or normal. The model processes individual images from the 3D CT scans and aggregates these image-level predictions to provide a comprehensive diagnosis for the entire volume. The results indicate that this deep learning approach achieves an impressive area under the curve (AUC) value of 96%, demonstrating its potential effectiveness in rapid and accurate detection of COVID-19 from CT imaging [15].

**Chapter 3**

# Dataset

In this study, we utilized a comprehensive dataset from **The Cancer Imaging Archive (TCIA)**, as illustrated in Fig. 1. The dataset specifically comprises the **LungCT-Diagnosis collection**, a well-curated repository of diagnostic, contrast-enhanced CT scans of lung adenocarcinoma patients. This dataset underpins the development and evaluation of our AI-driven diagnostic tools, facilitating in-depth analysis of lung tumors for improved prognostic accuracy.

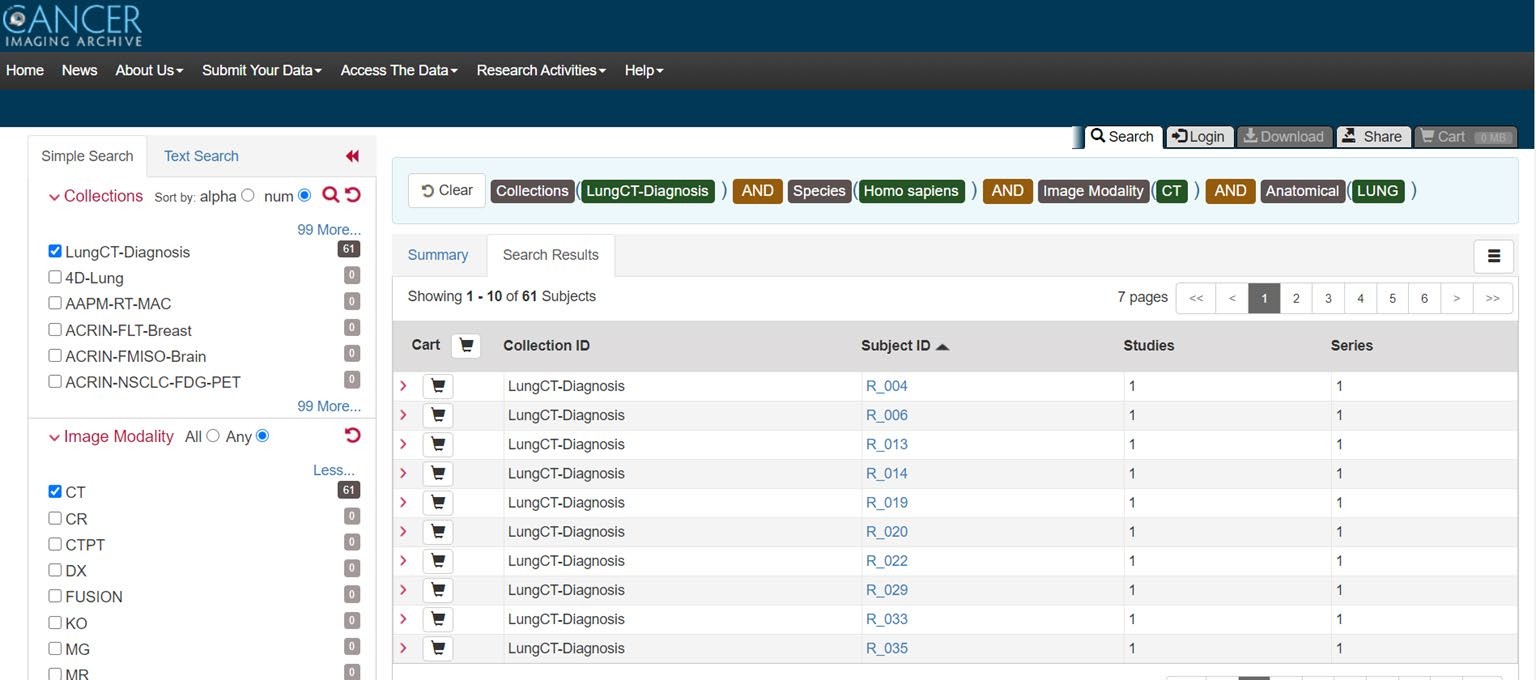


Figure 1 LungCT-Diagnosis dataset from Cancer Imaging Archive.

## DATA OVERVIEW

Title: Quantitative Computed Tomographic Descriptors Associate Tumor Shape Complexity and Intratumor Heterogeneity with Prognosis in Lung Adenocarcinoma

DOI: 10.7937/K9/TCIA.2015.A6V7JIWX

Location: Lung

Species: Human

Subjects: 61 patients

Data Types: CT images and supporting clinical data

Cancer Types: Lung Cancer

Size: 2.47 GB

Status: Public, Complete

This dataset serves as a valuable resource for extracting imaging biomarkers related to tumor morphology and density variation. As lung adenocarcinoma has unique structural and density characteristics that correlate with prognosis, this data enables the development of robust, quantitative tools for tumor analysis and survival prediction.

## DATASET SUMMARY

The **LungCT-Diagnosis collection** contains a total of **4,682 diagnostic contrast-enhanced CT images** (see Fig. 2) obtained retrospectively. Each CT scan was acquired with slice thicknesses between **3 mm and 6 mm**, standard for clinical imaging protocols. The images were retrospectively collected to ensure that patients had adequate follow-up before surgical intervention, allowing for a more accurate assessment of long-term survival outcomes.

The dataset's main purpose is to enable **prognostic feature extraction**, focusing on the identification and quantification of imaging markers that describe lung adenocarcinoma. Two essential quantitative features were developed:

1. **Tumor Shape Complexity**: Evaluates the irregularities in tumor contours and volume.
2. **Intratumor Density Variation**: Measures the fluctuations in density within the tumor, reflecting heterogeneity in cellular structure.

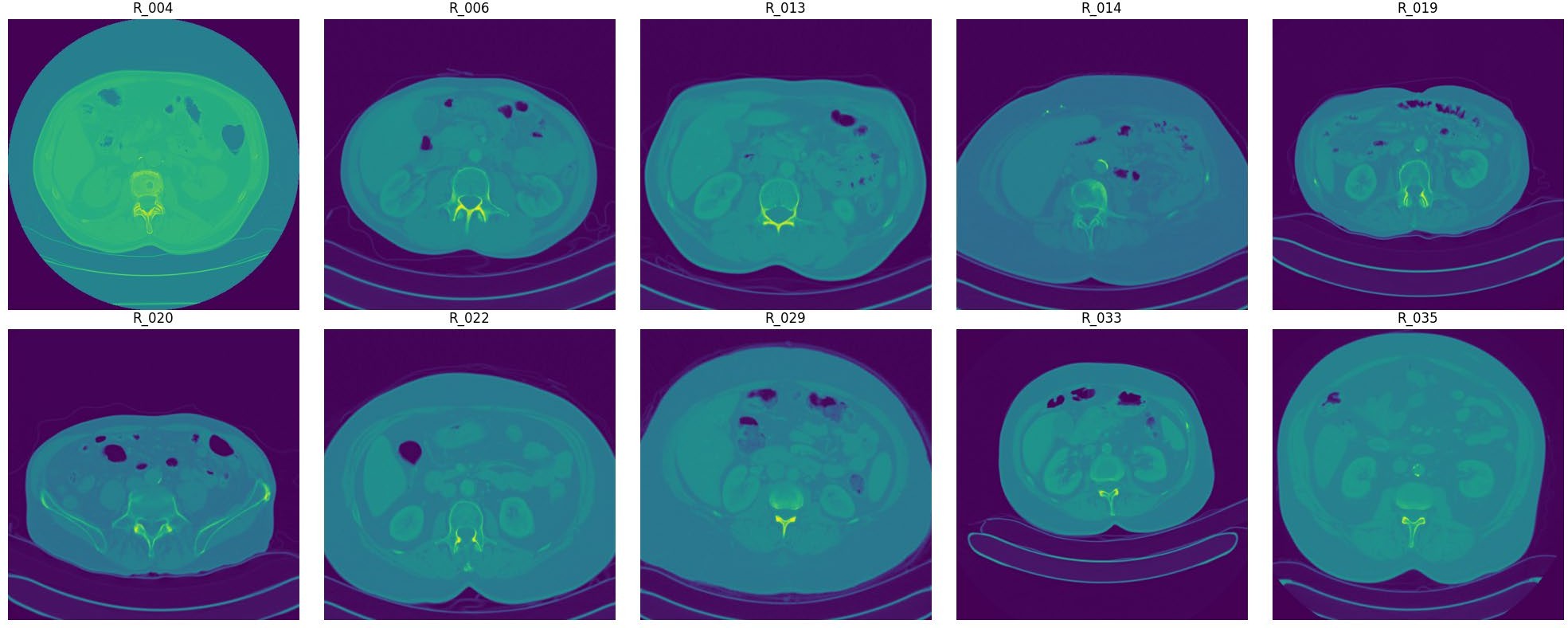
Both features are reproducible and show robustness across various clinical imaging protocols. Their stability despite differences in image acquisition parameters suggests that they could be integrated into routine diagnostic workflows for lung cancer management, providing valuable insights into patient outcomes.

Figure 2 Visualization of raw DICOM files of different subjects.

## DATA ACCESS

The **LungCT-Diagnosis dataset** is publicly accessible via the TCIA platform. The data includes:

* **CT Images**: Available in DICOM format (2.47 GB).
* **DICOM Metadata Digest**: Provided in CSV format for easy reference.
* **Representative Tumor Slices**: Offered in XLS format to streamline analysis.
* **Clinical Data**: Available in DOC format for comprehensive patient profiles.

Access to this data requires the **NBIA Data Retriever** tool, which facilitates the downloading of large DICOM datasets from TCIA. Researchers must adhere to the **TCIA Data Usage Policy**, ensuring appropriate citation and compliance with data-sharing protocols.

## DICOM FORMAT AND HOUNSFIELD UNITS

The DICOM (Digital Imaging and Communications in Medicine) format is the universal standard for storing, managing, and transferring medical imaging data, supporting interoperability across various imaging devices and software. Each DICOM file contains a series of cross-sectional slices of the patient’s anatomy, with metadata detailing patient demographics and imaging parameters.

Within DICOM files, Hounsfield Units (HU) are employed to represent tissue density values. Calculated based on X-ray attenuation, HU values allow clinicians to differentiate between various tissues and structures. For instance, tumors display unique HU patterns that can assist in identifying abnormal growths. This dataset uses the .dcm extension for DICOM files, organizing them within folders named according to patient IDs. Each folder contains multiple CT slices, offering a detailed, multi-angled view of lung structures.

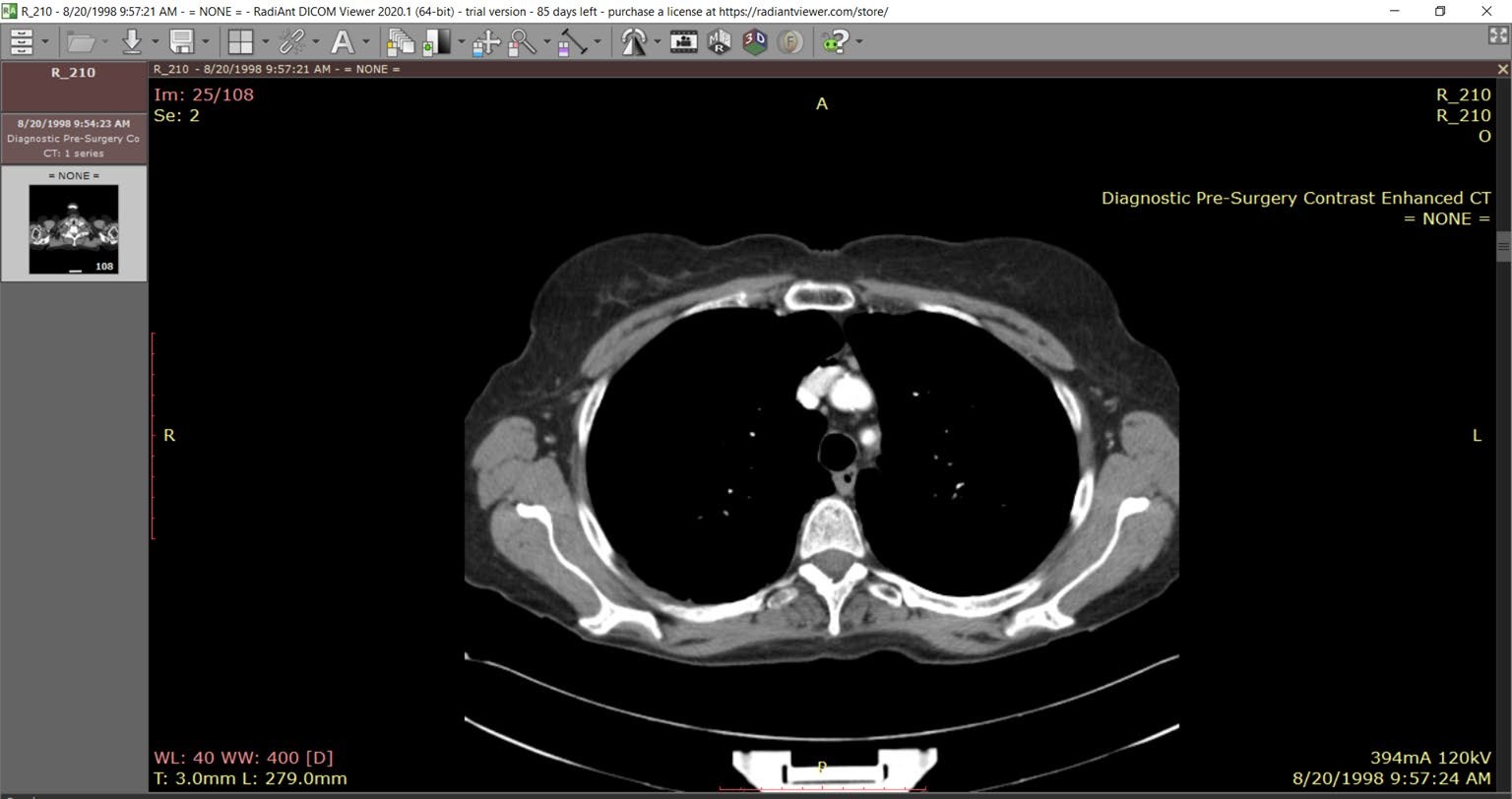


Figure 3 Visualization of a slice of CT scan of the subject R\_210 showing raw metadata attached to it.

To simplify analysis, patients are categorized based on survival status—either ALIVE (1) or DEAD (0)—providing a binary classification for outcome prediction. Fig. 3 illustrates a sample DICOM file structure, showing how these files enable precise, layered imaging across varied anatomical perspectives.

Due to variation in the number of slices per patient folder, diverse clinical scenarios are captured, reflecting real-world variability in imaging protocols. Tools such as QuPath and Radiant DICOM Viewer were utilized for visualization and analysis, enhancing our understanding of tumor morphology and potential pathologies present in the scans.

Figure 4 DICOM Slice visualized without DICOM Viewer using pydicom

## PREPROCESSING PIPELINE

To ensure data consistency and enhance model performance, a series of preprocessing steps were implemented:

* Normalization: Pixel intensity values were normalized across scans (see Fig. 5), creating a standardized range to facilitate consistent analysis. Normalization minimizes the effects of lighting and contrast variations, enabling more accurate feature extraction.
* Resampling: Voxel dimensions were resampled to 1mm x 1mm x 1mm, standardizing spatial resolution across all scans. This resampling ensures uniformity, which is essential for comparing tumors across different patients and imaging sessions.
* Segmentation: Using Hounsfield Units (HU), lung regions were segmented to isolate the relevant anatomical structures. By focusing solely on lung tissue, segmentation improves computational efficiency and accuracy in identifying tumors and analyzing their features.
* Augmentation: Data augmentation techniques, including rotation, flipping, and noise addition, were applied to diversify the training dataset (Fig. 6). These augmentations increase model robustness, helping the model generalize better to new, unseen data by simulating variations in tumor appearance and orientation.

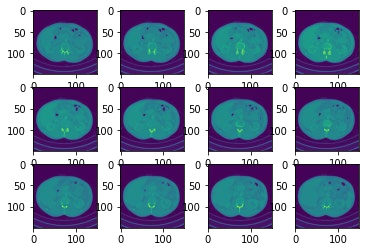


Figure 5 Data before preprocessing techniques.

## ANNOTATION AND LABELING

Each CT scan is meticulously annotated with labels to indicate the presence or absence of lung adenocarcinoma. The annotations also include critical clinical information such as **survival status** and **TNM staging** (tumor-node-metastasis classification).

Annotations were conducted by experienced radiologists, ensuring high-quality and reliable data labels for training AI models. This comprehensive labeling provides a solid foundation for supervised learning, where models can learn to associate imaging features with specific diagnostic and prognostic outcomes.

**Chapter 4**

# Methodology

The architecture of this project is organized into four primary components, each designed to handle specific stages within the diagnostic pipeline: **3D image processing**, **lung segmentation and visualization**, **tumor localization**, and **machine learning for diagnosis**. Each component applies distinct algorithms and frameworks to process medical images and contribute to a cohesive diagnostic model.

The pipeline starts by loading and preprocessing lung CT scans in **Digital Imaging and Communications in Medicine (DICOM)** format. These scans, consisting of multiple axial slices per patient, are processed using Python’s **pydicom library** to create a standardized 3D voxel grid for further analysis.

## 3D IMAGE PROCESSING

**Input**: Lung CT scans in DICOM format  
**Processing**: Sorting slices, converting pixel intensities to Hounsfield Units, and resampling to a consistent voxel size  
**Output**: Preprocessed 3D voxel grid for further analysis

The pipeline begins with the acquisition of lung CT scans, formatted in **DICOM (Digital Imaging and Communications in Medicine)**, which is the standard format for medical imaging data. Each CT scan comprises multiple axial slices for each patient. The **pydicom** library is employed to load and preprocess these DICOM files, ensuring they are organized and standardized for subsequent analysis.

**Slice Sorting and Hounsfield Unit Conversion**

1. **Sorting Slices**: Each DICOM slice is organized based on the ImagePositionPatient tag to maintain the correct anatomical sequence. This step preserves the natural spatial ordering of slices for accurate 3D reconstruction.
2. **Conversion to Hounsfield Units (HU)**: Each pixel intensity is converted to **Hounsfield Units** using the formula:

𝐻𝑈 = 𝑝𝑖𝑥𝑒𝑙\_𝑣𝑎𝑙𝑢𝑒 × 𝑅𝑒𝑠𝑐𝑎𝑙𝑒𝑆𝑙𝑜𝑝𝑒 + 𝑅𝑒𝑠𝑐𝑎𝑙𝑒𝐼𝑛𝑡𝑒𝑟𝑐𝑒𝑝𝑡

where RescaleSlope and RescaleIntercept are derived from the DICOM metadata. The HU scale represents tissue density, allowing the model to distinguish between air, lung tissue, and other structures.

**Resampling to Uniform Spacing**

To ensure consistency across scans, the images are resampled to an **isotropic resolution of 1mm x 1mm x 1mm** using cubic interpolation with scipy.ndimage.zoom. The new shape of each scan volume is calculated as follows:

𝑛𝑒𝑤\_𝑠ℎ𝑎𝑝𝑒 = (𝑛𝑒𝑤\_𝑠𝑝𝑎𝑐𝑖𝑛𝑔 / 𝑜𝑙𝑑\_𝑠𝑝𝑎𝑐𝑖𝑛𝑔) ∗ 𝑜𝑙𝑑\_𝑠ℎ𝑎𝑝𝑒

This resampling standardizes the spatial resolution, allowing uniform analysis across all 3D scans.

|  |
| --- |
| **Algorithm 1** Marching Cube Algorithm for 3D Lung  Visualisation |
| **Input:** |
| *scalarField:* 3D array of scalar values (e.g., lung data) |
| *isovalue:* scalar value to extract the surface |
| **Output:** |
| *triangles*: list of triangles representing the extracted surface |
| **Method:** |
| 1. Initialize *triangles* as an empty list. |
| 2. **For** each voxel in *scalarField*: |
| a. Calculate *cubeIndex* using *calculateCubeIndex(voxel,*  *isovalue).* |
| b. **For** each triangle in *getTriangleConfiguration(cubeIndex):* |
| i. **If** triangle is valid: |
| A. Interpolate vertices using  *interpolateTriangleVertices(voxel, triangle, isovalue).* |
| B. Add vertices to *triangles*. |
| 3. Return *triangles*. |

## 3D LUNG SEGMENTATION

**Input**: Preprocessed 3D voxel grid  
**Processing**: Thresholding, morphological operations, and region-growing  
**Output**: Binary mask of lungs, converted to a 3D mesh using Marching Cubes

The next phase focuses on segmenting the lung tissue, isolating it from other anatomical structures. Segmentation provides a focused area for tumor analysis and visualization.

|  |  |  |
| --- | --- | --- |
| **Tissue Type** | **Hounsfield Unit (HU)**  **Range** | **Description** |
| Air | -1005 to -995 | Lowest HU values, indicating the least dense material. |
| Lung | -950 to -550 | Low HU values, representing air-filled spaces in the lungs. |
| Fat | -100 to -80 | Relatively low HU values, indicating low tissue density. |
| Water | -4 to 4 | Neutral HU value, representing the density of water. |
| Kidney | 20 to 40 | Moderate HU values, suggesting a denser tissue  than water. |
| Pancreas | 30 to 50 | Slightly higher HU values compared to the kidney. |
| Blood | 50 to 60 | Moderately high HU values, indicating a denser tissue. |
| Muscle, Soft Tissue | 20 to 100 | A wide range of HU values, representing various soft tissues. |
| Liver | 50 to 70 | Relatively high HU values, indicating a denser tissue. |
| Adipose Tissue | -200 to -20 | Lower HU values compared to other tissues, representing fat. |
| Spongious Bone | 50 to 300 | Higher HU values, indicating a denser tissue than soft tissues. |
| Compact Bone (Cortical) | >300 | Highest HU values, representing the densest tissue in the body. |

**Thresholding and Morphological Operations**

Segmentation of lung tissue is based on **thresholding** in **Hounsfield Units (HU)**. Lung tissue typically falls within the HU range of **-950 to -550 HU**, while other tissues such as bone exhibit much higher values. Morphological operations are then applied to refine the lung mask:

* **Erosion and Dilation**: These operations remove small, isolated components and close gaps within the lung region.
* **Region-Growing**: This technique refines the segmented lung volume, specifically excluding structures like the heart and major blood vessels.

**3D Visualization using Marching Cubes**

Using the **Marching Cubes algorithm**, the binary lung mask is converted into a **3D surface mesh**. Marching Cubes identifies surface boundaries within the voxel grid, creating a triangulated mesh that visualizes the lung’s contours (see Table 1 for Hounsfield Unit thresholds by tissue type). This 3D visualization provides clinicians with an intuitive representation of lung structures.

* 1. 3D TUMOR LOCALIZATION (INTERACTIVE)
* Input: Segmented lung and tumor metadata
* Processing: Mapping tumor coordinates from real-world to voxel space, overlaying tumor on lung segmentation
* Output: 3D visualization of lung with tumor location marked

This stage maps the tumor’s location from clinical metadata onto the segmented 3D lung volume, allowing precise visualization of the tumor within the lung.

**Tumor Position Mapping**

Tumor coordinates, provided in real-world (millimeter) space, are transformed into voxel coordinates within the CT scan using the following equation:

𝑣𝑜𝑥𝑒𝑙\_𝑝𝑜𝑠𝑖𝑡𝑖𝑜𝑛 = (𝑟𝑒𝑎𝑙\_𝑤𝑜𝑟𝑙𝑑\_𝑝𝑜𝑠𝑖𝑡𝑖𝑜𝑛− 𝑜𝑟𝑖𝑔𝑖𝑛) / 𝑣𝑜𝑥𝑒𝑙\_𝑠𝑝𝑎𝑐𝑖𝑛𝑔

**Chapter 5**

# Methodology

The KITTI dataset is an essential resource for developing and testing computer vision algorithms in autonomous driving. The methodology for utilizing this dataset effectively involves multiple stages, from data preparation to model evaluation. Each step is carefully designed to ensure that the models can handle real-world driving challenges, from object detection to depth estimation and visual odometry. Below is an expanded breakdown of each stage in the methodology.

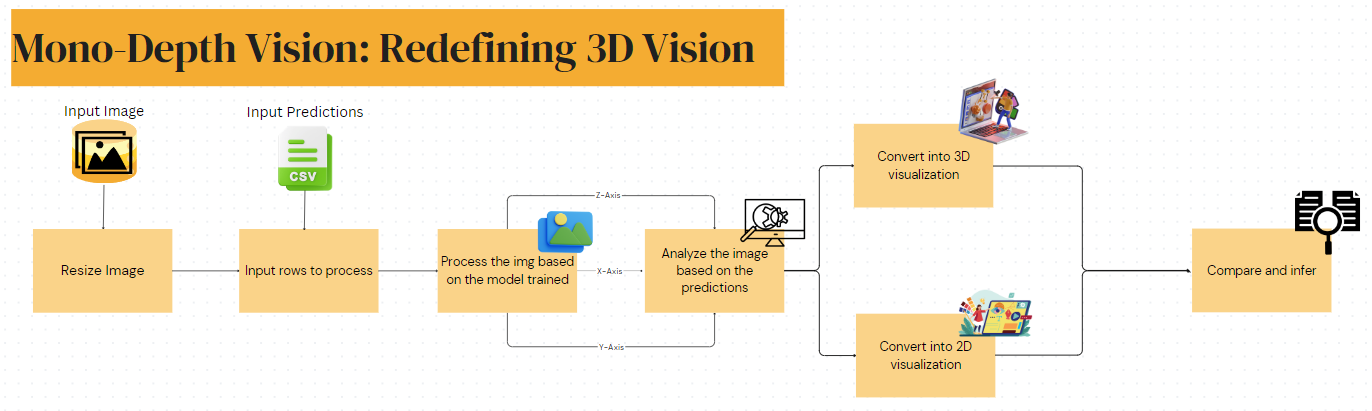


Figure 8: the flowchart of the methodology

## DATASET PREPARATION

The first phase in processing the KITTI dataset involves **downloading, organizing, and preparing** the data. The KITTI dataset contains various data types (e.g., stereo images, LiDAR point clouds, and annotations) that must be organized according to the requirements of the specific task, such as stereo vision, 3D object detection, or SLAM (Simultaneous Localization and Mapping).

5.1.1 DATA SELECTION

**Task-Specific Data**: Based on the research objective, only the relevant subsets of the KITTI dataset are selected. For example:

* For **stereo vision** and **depth estimation**, researchers primarily use the stereo image pairs (left and right images) to estimate depth from disparity.
* For **3D object detection**, the LiDAR point clouds and 3D bounding box annotations are essential for training models to localize objects in 3D space.
* For **SLAM** or **visual odometry** tasks, sequences of images, along with GPS and IMU data, are selected to help models estimate the vehicle’s trajectory and surroundings.

**Filtering and Segmentation**: If the task focuses on specific scenarios (e.g., urban vs. rural settings), the data may be filtered to include only relevant subsets. For instance, researchers working on urban navigation might choose images with dense traffic or complex urban environments.

5.1.2 DATA SPLIT

* **Pre-Defined Splits:** The KITTI dataset provides predefined training and testing splits for each task to ensure consistent evaluation across models.
* **Custom Splits for Validation:** In many cases, researchers further divide the training set into a training subset and a validation subset. For example, an 80-20 split of the training data is often used to create a validation set for tuning model parameters, preventing overfitting and ensuring generalizability.
* **Cross-Validation:** Some tasks benefit from cross-validation, especially when data is limited. Cross-validation techniques, like k-fold validation, allow models to be evaluated on multiple subsets, improving robustness by averaging performance across different validation sets.

5.1.3 DATA AUGUMENTATION

Data augmentation is applied to expand the effective size of the training dataset, which helps improve model generalization. Common augmentation techniques include:

* **Random Cropping**: Crops are applied randomly to simulate partial views of objects or scenes, helping models generalize across different scales and viewpoints.
* **Flipping**: Horizontal flipping is frequently used, especially for images, to make models invariant to left-right orientation, which can be useful in scenarios with symmetric features.
* **Scaling**: Scaling the images up or down can help models learn to detect objects at various distances.
* **Color Jittering**: Adjusting the brightness, contrast, saturation, or hue can simulate variations in lighting conditions, making the model robust to different environments.
* **Gaussian Noise**: Adding random noise, especially to LiDAR point clouds, can make models more resilient to sensor noise, which is common in real-world autonomous driving scenarios.

## PREPROCESSING

Preprocessing is essential for ensuring that the data is in the correct format and ready for model training. This stage includes calibration, alignment, parsing, and normalization.

5.2.1 CALIBERATION AND RECTIFICATION

* **Intrinsic and Extrinsic Calibration**: KITTI provides intrinsic (camera-specific) and extrinsic (relative to other sensors) calibration parameters for each camera and the LiDAR sensor. These parameters correct lens distortion and align the stereo images, enabling accurate depth estimation from disparity maps.
* **Rectification of Stereo Images**: Rectification aligns the left and right images so that corresponding points in both images appear on the same horizontal line. This is crucial for stereo vision, as it allows disparity to be calculated along the horizontal axis.
* **LiDAR and Camera Alignment**: For tasks combining LiDAR and image data, the LiDAR point clouds are aligned with the camera frame using calibration data. This alignment is important for multi-sensor fusion, allowing the model to merge information from both 2D images and 3D point clouds.

5.2.2 ANNOTATION PARSING

For tasks that involve object detection or tracking, KITTI provides annotation files with bounding box coordinates, object classes, and occlusion/truncation information. Parsing these files is necessary to extract and organize the annotations:

* **2D and 3D Bounding Boxes**: For 2D detection, bounding box coordinates are extracted for each object. For 3D detection, additional dimensions (height, width, length) and rotation angles are parsed.
* **Object Labels**: Classes like cars, pedestrians, and cyclists are parsed from annotations to train object-specific detection models.
* **Occlusion and Truncation Information**: This data indicates how much of an object is obscured or cut off by the image border, helping models account for partially visible objects.

5.2.3 DATA NORMALIZATION

* **Image Normalization**: Pixel values are typically scaled to a specific range, such as [0, 1] or standardized to zero mean and unit variance. This improves convergence speed during training.
* **Point Cloud Normalization**: For 3D data, point clouds may be normalized through transformations like scaling and translation to fit within a consistent range. Normalized point clouds make it easier for models to generalize across scenes with varying distances and scales.

## MODEL DESIGN

Based on the specific task, researchers select or design model architectures tailored to process the KITTI dataset effectively. Below are the common architectures used for different tasks.

5.3.1 STEREO VISION AND DEPTH ESTIMATION

* **Convolutional Neural Networks (CNNs):** CNNs are often used for image processing tasks like monocular depth estimation, taking single images as input to predict depth maps.
* **U-Net and Hourglass Networks:** These architectures are popular for dense prediction tasks, such as disparity and depth estimation. They are designed to capture both low-level and high-level features, making them suitable for pixel-wise predictions.
* **Cost Volume Networks:** For stereo depth estimation, networks may use a "cost volume" to compare features across the left and right images, identifying corresponding points for depth calculation.



Figure 9- image/target/prediction for the first epoch

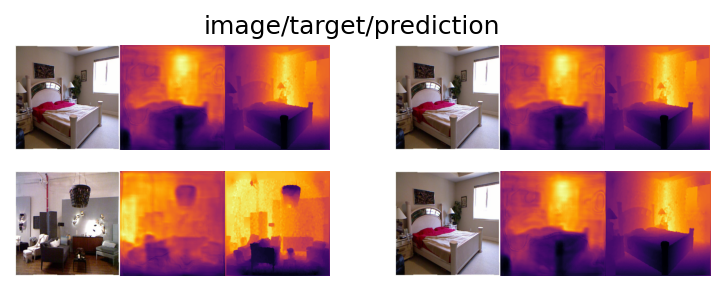


Figure 10- image/target/prediction for the best epoch

5.3.2 2D AND 3D OBJECT DETECTION

* **YOLO and Faster R-CNN:** These architectures are used for 2D detection tasks, with YOLO known for speed and Faster R-CNN for accuracy. They process 2D images to predict bounding boxes and class labels.
* **VoxelNet, PointPillars, and 3D SSD:** For 3D object detection, these architectures process point clouds directly. VoxelNet and PointPillars convert point clouds into grid representations, while 3D SSD operates on raw points, generating 3D bounding boxes for each detected object.

5.3.3 VISUAL ODOMETRY AND SLAM

**Recurrent Neural Networks (RNNs)**: RNNs are used in visual odometry tasks for sequential data processing, capturing temporal information to predict vehicle motion.

**CNNs with Bundle Adjustment**: CNNs are combined with optimization techniques like bundle adjustment to estimate motion by refining multi-frame depth maps.

**DeepVO and PoseNet**: These architectures are commonly used for end-to-end visual odometry, where images are input to directly estimate the vehicle’s trajectory.

5.3.4 SENSORY FUSION

In sensor fusion, data from multiple sensors is combined to enhance perception:

* **Early Fusion**: Involves combining raw sensor data (e.g., concatenating stereo images with point clouds) before inputting it to the model.
* **Late Fusion**: Features extracted from each sensor are fused at a later stage, such as merging depth and object detection features for a unified 3D scene representation.

## TRAINING AND HYPERPARAMETER TUNING

Training involves optimizing the model parameters to minimize errors, while hyperparameter tuning adjusts settings for optimal performance.

5.4.1 LOSS FUNCTION

**Depth Estimation**: Common loss functions include mean squared error (MSE) and L1 loss, which measure the pixel-wise difference between predicted and ground-truth depth.

**Object Detection**: Object detection loss functions often combine classification loss (e.g., cross-entropy for class prediction) and regression loss (e.g., smooth L1 for bounding box localization).

**Visual Odometry**: Pose estimation loss, such as the Euclidean distance between predicted and ground-truth trajectories, is used to train visual odometry models.

5.4.2 OPTIMZATION

**Gradient-Based Methods**: Optimizers like Stochastic Gradient Descent (SGD) and Adam are widely used. These methods iteratively adjust weights to minimize the loss function.

**Learning Rate Scheduling**: Learning rate schedules reduce the learning rate during training, often improving convergence. Techniques like cosine annealing or exponential decay are common.

**Early Stopping**: Early stopping halts training when validation performance plateaus, preventing overfitting.

5.4.3 HYPERPARAMETER TUNING

**Grid Search**: Systematic exploration of hyperparameter combinations, such as batch size, learning rate, and regularization strength.



Figure 11- Image/target/prediction of the first 10 images using the trained model

**Bayesian Optimization**: Bayesian optimization can be more efficient, as it uses probabilistic models to explore the hyperparameter space effectively.

## EVALUATION

After training, the model’s performance is evaluated using metrics and benchmarks.

5.5.1 METRICS

**2D Object Detection**: Mean Average Precision (mAP) measures the average precision across all object classes and difficulty levels.

**Depth Estimation**: Metrics like root mean square error (RMSE) and absolute relative error (Abs Rel) evaluate the accuracy of predicted depth maps.

**Optical Flow**: End-point error (EPE) calculates the deviation between predicted and ground-truth flow vectors.

**3D Object Detection**: 3D Intersection over Union (IoU) and average precision (AP) evaluate detection accuracy in 3D space.

5.5.2 BENCHMARKING

**KITTI Evaluation Server**: For consistency, results on the test set are submitted to the KITTI server. The server provides ranking and scores, ensuring fair comparison across models.

## POST PROCESSING AND VIZUALIZATION

Post-processing serves to refine, interpret, and visualize the predictions made by models trained on the KITTI dataset. This phase involves several key steps to improve model output quality and to better understand model performance.

5.6.1 NON-MAXIMUM SUPRESSION (NMS) IN OBJECT DETECTION

In object detection tasks, models often produce multiple overlapping bounding boxes for the same object. **Non-Maximum Suppression (NMS)** is a post-processing technique used to remove redundant bounding boxes, ensuring that only the most accurate prediction is retained for each detected object.

* **Why NMS is Needed**: During object detection, models can generate multiple bounding boxes around a single object, especially when it appears in challenging positions (e.g., occluded, partially visible). Overlapping boxes often correspond to the same object but differ in location, size, or confidence score, leading to redundancy.
* **How NMS Works**:
  + **Confidence Thresholding**: NMS begins by discarding any bounding boxes with a confidence score below a specified threshold. This removes weak detections, focusing only on the boxes most likely to contain an object.
  + **IoU Calculation**: For each pair of remaining bounding boxes, NMS calculates the **Intersection over Union (IoU)**, a metric that quantifies the overlap between boxes.
  + **Selecting the Maximum**: Starting with the bounding box that has the highest confidence score, NMS removes any overlapping boxes with an IoU greater than a predefined threshold (typically around 0.5). This process repeats until only one bounding box per object remains.
* **Adaptive NMS Variants**: Variants of NMS, like **soft NMS** (which reduces confidence scores of overlapping boxes rather than removing them) and **class-aware NMS** (which applies NMS separately for each object class), are sometimes used to further improve detection accuracy and precision.

NMS is essential for producing clean, non-overlapping bounding boxes, which significantly improves the readability of the detection results, especially in scenes with high object density, such as urban areas in KITTI.

5.6.2 DEPTH ESTIMATION REFINEMENT

For depth estimation tasks, post-processing aims to smooth the predicted depth maps and remove artifacts. Predicted depth maps may contain noise, especially at object boundaries and in areas with complex textures or occlusions. Common techniques for depth refinement include:

* **Bilateral Filtering**:
  + Bilateral filtering is a smoothing technique that preserves edges in depth maps while removing noise in homogeneous regions.
  + This filter smooths depth values by considering both spatial proximity and pixel similarity, meaning that it applies stronger smoothing in uniform areas while preserving boundaries where abrupt depth changes occur.
  + Bilateral filtering is particularly useful in automotive scenes where maintaining sharp object boundaries (e.g., between cars, pedestrians, and the road) is crucial for accurate perception.
* **Depth Map Denoising**:
  + Denoising techniques, such as **median filtering** or **Gaussian smoothing**, are applied to reduce noise caused by imperfect predictions. Median filtering helps remove outliers without heavily blurring boundaries.
  + These techniques help to reduce speckled noise in depth maps, particularly in areas where sensors struggle (e.g., glossy or transparent surfaces).
* **Post-Processing Networks**:
  + In some cases, researchers apply an additional neural network specifically designed for post-processing depth maps. This network can learn to identify and correct common artifacts and improve depth accuracy at object boundaries.
  + Post-processing networks often use high-resolution input and output layers, fine-tuning the depth predictions from the main model to increase clarity and realism.

Depth map refinement is especially important in autonomous driving, where accurate and smooth depth estimates are essential for tasks like obstacle avoidance and navigation in dynamic environments.

5.6.3 VISUALIZATION OF MODEL OUTPUTS

Visualization is an integral part of post-processing, enabling researchers to assess model performance qualitatively. It helps in identifying errors, understanding limitations, and communicating results effectively. Here are some common visualization techniques used with KITTI data:

* **3D Bounding Box Visualization**:
  + For 3D object detection tasks, visualizing 3D bounding boxes overlaid on point clouds or images helps validate the spatial accuracy of detected objects.
  + Bounding boxes can be color-coded based on class (e.g., green for cars, blue for pedestrians) or confidence score to distinguish objects at a glance.
  + In complex scenes, overlaying 3D bounding boxes on top of LiDAR point clouds or images provides a comprehensive view of object locations in 3D space, helping validate the accuracy of predictions.
* **Depth Map Visualization**:
  + Depth maps are visualized using color gradients, where colors represent different depth values. For example, close objects might be colored red, while distant objects are blue.
  + This visualization allows for quick assessment of depth estimation quality, particularly at object boundaries and in occluded regions.
  + Overlaying depth maps on the original images is also common, as it provides a direct comparison between the estimated depths and the visual scene.
* **Optical Flow Visualization**:
  + Optical flow fields, which represent pixel-wise motion between consecutive frames, are visualized as color-coded vectors or gradient fields.
  + Color coding helps indicate the direction and magnitude of motion, providing insights into how the model interprets the movement of objects and the vehicle’s own motion.
  + This visualization is particularly useful for tasks like visual odometry, where accurate flow predictions are essential for trajectory estimation.

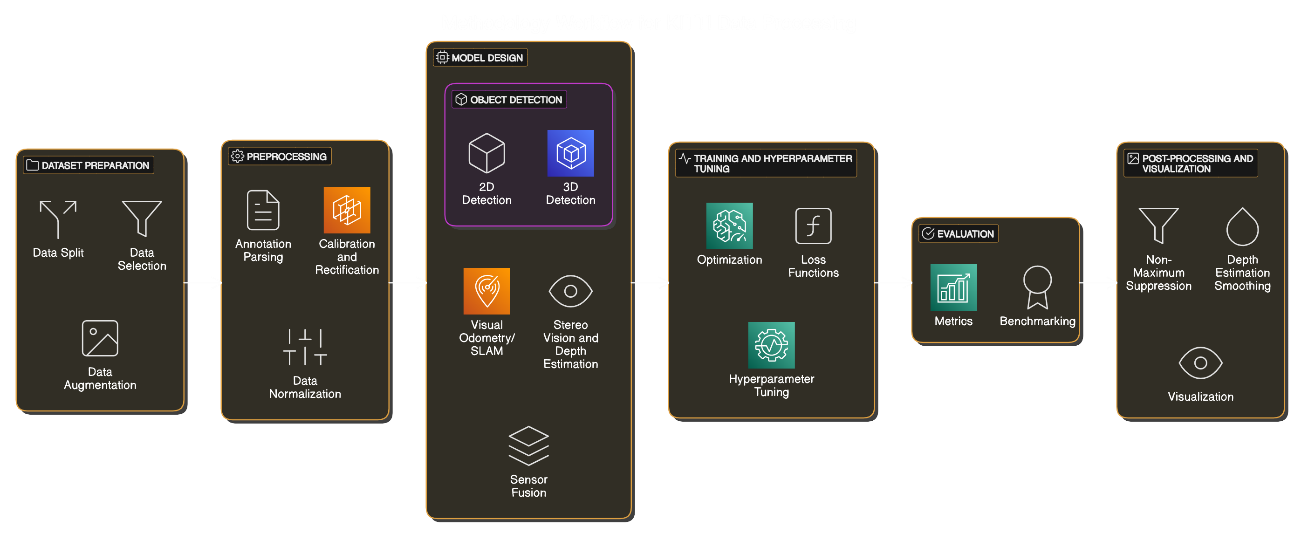


Figure 12: Describes the entire methodology in one flowchart

**Chapter 6**

# 2D Training

Monocular depth estimation, the task of predicting depth from a single 2D image, is a fundamental challenge in computer vision. It has numerous applications in fields such as autonomous driving, augmented reality, and robotics. Traditionally, depth estimation required stereo vision or multiple viewpoints. However, advancements in deep learning, particularly with convolutional neural networks (CNNs), have enabled the accurate prediction of depth from single images. This shift reduces hardware complexity and makes depth estimation more accessible, allowing models to infer 3D scene structures from only a single RGB image. Here is a detailed methodology for training and optimizing monocular depth estimation models.

## TRAINING DATASET

The quality and diversity of training data are crucial for the success of monocular depth estimation models. High-quality, large-scale datasets with paired RGB and depth images allow the model to learn relationships between image features and depth.



Figure 13: 2D Training workflow

6.1.1 KITTI DEPTH V2 DATASET

The **KITTI Depth V2 dataset** is commonly used in monocular depth estimation because of its diverse scenarios and high-quality annotations. This dataset was collected in real-world driving conditions, offering a range of environmental settings, object types, and distances. Key features of the KITTI Depth V2 dataset include:

* **Diverse Environments**: The dataset includes images from urban, rural, and highway settings, providing a wide variety of scenes for the model to learn from.
* **High-Resolution Annotations**: Each RGB image is paired with a high-resolution depth map derived from LiDAR data, which serves as the ground truth for training.
* **Variety of Conditions**: KITTI captures images under different lighting and weather conditions, making it ideal for training robust models that can generalize across various real-world situations.

6.1.2 DATASET SPLIT

To facilitate robust training, the KITTI dataset is typically divided into **training, validation, and testing sets**. These splits allow the model to be evaluated on data it hasn’t seen during training, providing insight into its generalization capabilities:

* **Training Set**: The majority of the data (e.g., 70-80%) is used for training the model. This set allows the model to learn feature-depth relationships.
* **Validation Set**: A small portion (e.g., 10-15%) is reserved for validation, helping to fine-tune hyperparameters and assess the model's generalization during training.
* **Testing Set**: The remaining data is used to evaluate the final model’s performance. This set includes images that the model has not encountered, providing an unbiased evaluation of its effectiveness.

## DATA PROCESSING

Data preprocessing is a critical step that ensures input images are in the right format for the neural network. Proper preprocessing improves model stability, speeds up training, and enhances generalization.

6.2.1 RESIZING

Images are typically resized to match the input dimensions expected by the model:

* **Consistency in Dimensions**: The KITTI dataset images are resized to a fixed input size, such as 256 × 512 pixels or 384 × 768 pixels, depending on the model’s architecture.
* **Preserving Aspect Ratio**: In some cases, resizing maintains the original aspect ratio to prevent image distortion, ensuring that spatial relationships between objects are preserved.

6.2.2 NORMALIZATION

Normalization helps the model converge faster and prevents numerical instability:

* **Pixel Value Scaling**: Pixel values are scaled to a range of [0, 1] or [-1, 1]. Scaling reduces the variation in input data, making it easier for the network to process.
* **Mean-Variance Normalization**: Some models use mean-variance normalization, where the mean pixel value is subtracted, and the result is divided by the standard deviation. This normalization aligns the input data distribution with the initialization of the model weights.

6.2.3 DATA AUGUMENTATION

Data augmentation artificially increases the training data diversity by applying transformations to the images:

* **Random Rotations**: Small rotations help the model learn rotational invariance, improving its ability to predict depth accurately regardless of orientation.
* **Adaptive Flipping**: Horizontal flips simulate the appearance of objects from different angles, which is particularly useful for autonomous driving applications.
* **Cropping and Scaling**: Random cropping and scaling help the model learn to recognize objects and depth relationships at different sizes and positions within the frame.
* **Color Jittering**: Adjusting brightness, contrast, and saturation enhances the model’s robustness to varying lighting conditions, which is critical for real-world applications like self-driving cars.
* **Occlusion Simulation**: Some augmentation techniques simulate occlusions by randomly blocking parts of the image, helping the model learn to infer depth even when objects are partially obstructed.

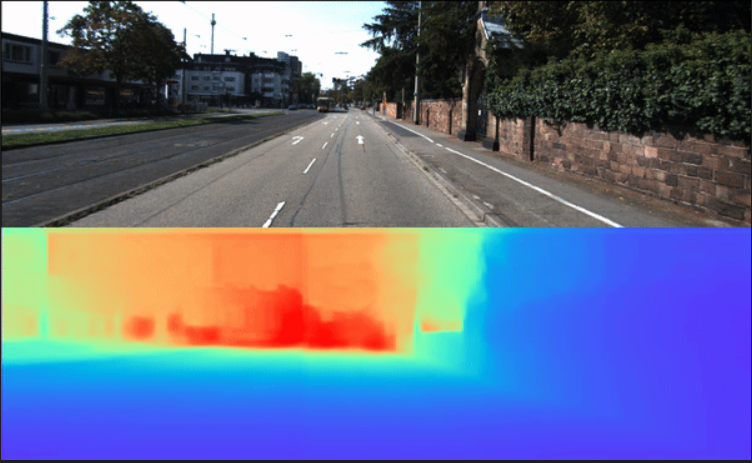


Figure 14: Augmenting the given dataset

## MODEL ARCHITECTURE

The architecture of the model is designed to capture spatial features and predict depth accurately. Different architectures are chosen based on their ability to generalize well, handle complex image features, and efficiently process high-resolution data.

6.3.1 CONVULTIONAL NEURAL NETWORKS(CNN)

CNNs are the backbone of most depth estimation models. They are effective at capturing spatial hierarchies in images, from low-level edges to high-level object features:

* **Encoder-Decoder Architecture**: Many depth estimation models use an encoder-decoder structure. The encoder extracts features from the input image, while the decoder maps these features back to the pixel-level depth map.
* **U-Net and Hourglass Networks**: These architectures are popular for dense prediction tasks. U-Net uses skip connections to preserve spatial information across layers, while hourglass networks process images at multiple scales to capture fine details and global context.
* **Pyramid Networks**: Pyramid-based networks process the image at multiple resolutions, capturing both fine-grained and coarse depth information.

6.3.2 HANDLING SCALE INVARIANCE

Monocular depth estimation faces challenges with objects of varying scales. Models like **MiDaS** use scale-invariant loss functions that focus on preserving relative rather than absolute depth:

* **Scale-Invariant Loss**: This loss function minimizes the relative differences between pixels, allowing the model to predict depth accurately regardless of absolute scale.
* **Multi-Scale Feature Extraction**: Some architectures extract features at multiple scales and incorporate them into the depth prediction, improving accuracy for objects at different distances.



Figure 15: Multi-scale feature Extraction

6.3.3 REGULARIZATION TECHNIQUES

To prevent overfitting, regularization methods are applied:

* **L2 Regularization**: Adds a penalty proportional to the sum of squared weights, encouraging simpler models that generalize better.
* **Early Stopping**: Monitors performance on the validation set and stops training when performance plateaus, preventing over-optimization on the training set.
* **Dropout**: Randomly "drops" neurons during training, which prevents the model from relying too heavily on any one feature and improves generalization.

## LOSS FUNCTION

The loss function is central to guiding the model's learning process by quantifying the error between predicted and actual depth values. Effective loss functions improve prediction accuracy and enable the model to learn meaningful depth representations.

6.4.1 DISPARITY MINIMIZATION

In monocular depth estimation, the objective is to minimize **disparity** or pixel-wise differences between the predicted depth map and the ground truth:

* **Mean Squared Error (MSE)**: The model calculates the squared difference between predicted and actual depth values for each pixel and aggregates these differences. MSE penalizes large errors more than small ones, encouraging the model to focus on substantial discrepancies.
* **L1 Loss**: This loss calculates the absolute difference between predicted and true depth values, often making it more robust to outliers than MSE.

6.4.2 HANDLING SCALE INVARIANCE

Scale-invariance is crucial in monocular depth estimation because objects at different distances may appear larger or smaller. To address this:

* **Relative Depth Loss**: Instead of minimizing absolute depth errors, relative depth loss minimizes errors in the relative differences between neighboring pixels, allowing the model to focus on depth gradients rather than absolute depth.
* **Logarithmic Loss**: Logarithmic losses can reduce the impact of depth errors in faraway objects, emphasizing accuracy in the immediate vicinity of the camera where depth perception is more critical.

6.4.3 REGULARIZATION IN LOSS

Regularization in the loss function helps avoid overfitting by adding penalties to the model's predictions:

* **Edge-Aware Loss**: This regularization technique penalizes errors in depth discontinuities, preserving edges and object boundaries.
* **Smoothness Constraints**: Depth maps often include smoothness constraints, encouraging neighboring pixels to have similar depths when they belong to the same object. This regularization enhances spatial coherence in the predicted depth maps.

## OPTIMZATION TECHNIQUES

Optimizing the training process ensures that the model converges efficiently, achieving accurate predictions without overfitting.

6.5.1 DYNAMIC LEARNING RATE DECAY

The learning rate, which controls how quickly the model updates its weights, may decay dynamically during training:

* **Scheduled Decay**: In scheduled decay, the learning rate decreases at predefined intervals or based on training epochs. This enables the model to make larger adjustments initially and smaller ones as it approaches an optimal solution.
* **Adaptive Decay**: Adaptive methods, such as the **ReduceLROnPlateau** scheduler, monitor the validation loss and reduce the learning rate if performance stagnates.

6.5.2 EARLY STOPPING

Early stopping prevents the model from overfitting by monitoring the validation loss:

* **Validation Monitoring**: The model’s performance on the validation set is checked after each epoch. If the performance does not improve after a certain number of epochs, training is halted to avoid overfitting.
* **Patience Parameter**: The patience parameter specifies how long the model should wait before stopping. A higher patience allows for more exploration, while a lower patience prevents overtraining.

6.5.3 BATCH NORMALIZATION AND DROPOUT

These techniques enhance stability and prevent overfitting:

* **Batch Normalization**: Normalizes the output of each layer, reducing internal covariate shifts and enabling the model to converge faster.
* **Dropout**: During training, a percentage of neurons are randomly disabled in each layer, preventing the model from becoming too dependent on specific neurons. This increases the model’s robustness.

## EVALUATION METRICS AND BENCHMARKING

After training, the model's performance is evaluated using specific metrics that quantify the accuracy of depth predictions. Benchmarking involves comparing the model’s performance against standard datasets.

6.6.1 KEY-EVALUATION METRICS

**Root Mean Squared Error (RMSE)**: Measures the average squared difference between predicted and actual depths, penalizing larger errors.

**Mean Absolute Error (MAE)**: Calculates the average absolute difference, providing a more robust measure that is less sensitive to outliers.

**Scale-Invariant Logarithmic Error**: Useful for monocular depth estimation, as it focuses on relative depth accuracy rather than absolute scale.

6.6.2 BENCHMARKING ON KITTI DATASET

The KITTI evaluation server allows researchers to compare their models against others on a standardized test set:

* **Public Leaderboards**: KITTI’s leaderboard ranks models based on evaluation metrics, providing a fair comparison of model performance.
* **Test Set Submission**: After final training, predictions are submitted to the server, which evaluates them against hidden ground-truth data for unbiased scoring.

## POST PROCESSING AND VISUALIZATIONS

Post-processing refines the depth predictions and prepares them for analysis, while visualization allows researchers to interpret model outputs qualitatively.

6.7.1 SMOOTHING AND REFINEMENT

**Bilateral Filtering**: Smooths depth maps while preserving edges, improving depth continuity within objects and enhancing boundaries.

**Depth Map Refinement**: Post-processing networks may be used to reduce artifacts and sharpen the depth map.

6.7.2 VISUALIZATION

**Color Mapping**: Depth maps are visualized with color gradients, where warmer colors (e.g., red) represent closer objects and cooler colors (e.g., blue) indicate farther objects.

**Overlaying on Original Images**: Depth predictions can be overlaid on the original image, allowing for a qualitative assessment of the model’s performance.

**Chapter 7**

# 3D Training

3D depth estimation is the process of predicting the depth and spatial relationships within a scene from 2D images, extending beyond the pixel-wise depth predictions in 2D to encompass 3D geometry and structure. This task is crucial for autonomous systems like self-driving cars, robotics, and augmented reality, where an accurate understanding of the 3D world enables more reliable interaction with the environment. Here’s a step-by-step breakdown of the training process for 3D depth estimation.

## 3D DATA AND DATASET

3D depth estimation requires high-quality 3D data that captures real-world spatial structures and relationships. Datasets for 3D training usually include paired 2D images with their 3D depth counterparts, such as point clouds or voxel grids, allowing the model to learn to infer depth from visual cues.

7.1.1 3D GROUND TRUTH

For training in 3D depth estimation, datasets need to provide accurate 3D ground truth:

* **Point Clouds**: Datasets like KITTI and NYU Depth V2 offer 3D point clouds generated by LiDAR sensors, which provide precise measurements of distances between the camera and objects in the scene. These point clouds serve as ground truth for the model to learn accurate depth predictions.
* **Stereo Image Pairs with Depth Information**: Stereo image pairs allow models to compute depth from disparity between two viewpoints. The depth derived from these pairs, either through stereo vision algorithms or from LiDAR-generated ground truth, enables the model to learn spatial relationships from monocular images.

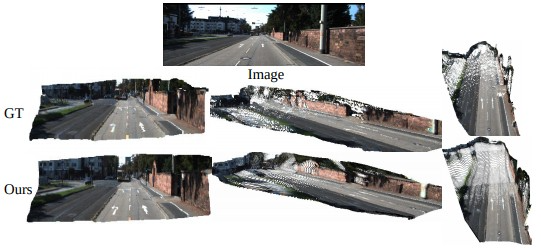


Figure 16: Image pairing

7.1.2 PAIRED 2D AND 2D DATA

Paired 2D and 3D data facilitate the learning process, allowing the model to map features in a 2D image to their corresponding 3D structures:

* **Mapping 2D Pixels to 3D Coordinates**: Training data that includes both 2D RGB images and corresponding 3D data, like point clouds or voxel grids, helps the model associate visual features in the 2D image with real-world 3D coordinates.
* **Rich Depth Annotations**: Datasets that provide high-quality annotations enable the model to generalize well to complex, real-world environments, as it learns to predict depth in varied conditions such as different lighting, weather, and occlusion levels.

## MODEL ARCHITECTURE

To handle the added complexities of 3D data, 3D depth estimation models incorporate specialized architectures that process spatial and depth information simultaneously, extending the principles of 2D CNNs.

7.2.1 3D CONVOLUTIONAL NEURAL NETWORKS(3D-CNN)

3D-CNNs are an extension of standard CNNs, designed to work with 3D data:

* **Volumetric Data Processing**: Unlike 2D-CNNs, which analyze width and height, 3D-CNNs also process depth, allowing them to understand volumetric features in 3D data.
* **3D Feature Extraction**: By analyzing voxels (3D pixels), 3D-CNNs can capture the structure and spatial relationships within the 3D space, making them well-suited for tasks like 3D object localization and depth estimation.

7.2.2 VOXEL REPRESENTATION

A common approach for handling 3D data is to convert it into a voxel grid, which represents the scene as a collection of small 3D units:

* **Voxel Grid Conversion**: Point clouds or 3D point data are often converted into voxel grids, where each voxel represents a small volume in 3D space. This representation enables the model to assess the "occupancy" of each voxel, effectively reconstructing the 3D scene.
* **Voxel Occupancy Prediction**: In depth estimation, the model predicts which voxels are occupied by objects and the relative depth of each, creating a 3D reconstruction aligned with real-world depth.



Figure 17: Voxel Representation

7.2.3 FEATURE FUSION

Feature fusion is a technique where 2D and 3D features are combined to improve depth prediction:

* **2D-3D Feature Combination**: Features extracted from 2D images (e.g., textures and edges) are fused with features from 3D data (e.g., spatial depth). This allows the model to leverage both types of information to make more accurate predictions.
* **Multi-Modal Input Integration**: Some models are designed to handle multiple types of data simultaneously (e.g., RGB images and LiDAR point clouds), merging information from different modalities to create a comprehensive depth map.

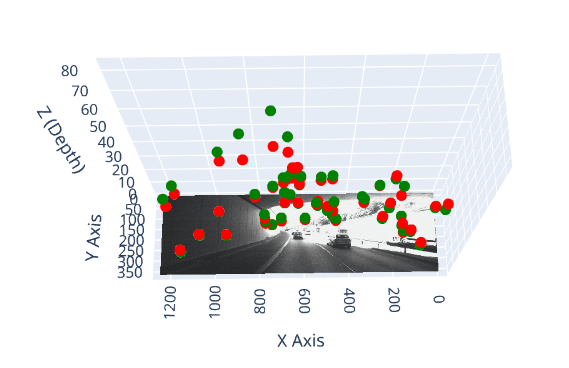


Figure 18: Integration of 2D and 3D

## LOSS FUNCTION

The loss function in 3D depth estimation is more intricate than in 2D depth estimation, as it must account for 3D geometric relationships and depth accuracy across three dimensions.

7.3.1 3D-DISPARITY LOSS

3D disparity loss minimizes the difference between predicted and ground truth 3D structures:

* **Point-to-Point Distance**: This loss calculates the Euclidean distance between predicted and actual 3D points. By minimizing this distance, the model learns to position objects accurately in 3D space.
* **Point-to-Plane Distance**: An alternative to point-to-point loss, point-to-plane loss calculates the distance from each predicted point to a plane fitted to nearby ground truth points. This approach reduces surface errors and aligns predicted points more closely with true surfaces.

7.3.2 GEOMETRIC CONSISTENCY LOSS

Geometric consistency loss enforces accurate spatial relationships within the predicted 3D structure:

* **Spatial Relationship Constraints**: This loss penalizes the model if predicted depth violates geometric constraints, such as overlapping objects or incorrectly ordered depth (e.g., predicting a car behind a tree as closer to the camera).
* **Physically Realistic Predictions**: This loss enforces depth predictions that respect physical constraints, ensuring the 3D scene remains coherent and realistic.

7.3.3 SURFACE NORMAL LOSS

Surface normal loss ensures that predicted 3D surfaces align correctly with real-world object boundaries and orientations:

* **Surface Normal Alignment**: The model predicts the orientation of surfaces in 3D, penalizing deviations from the ground truth surface normals.
* **Improved Boundary Precision**: By learning accurate surface normals, the model can better represent object boundaries and details, leading to more realistic depth predictions.

7.3.4 PHOTOMETRIC LOSS

When ground truth 3D data is unavailable, photometric loss can be used for unsupervised training:

* **Frame Synthesis**: In video sequences, the model predicts 3D depth for a frame, then synthesizes the next frame using the predicted depth map. The photometric loss compares the synthesized frame with the actual next frame, updating the model to reduce discrepancies.
* **Color and Intensity Matching**: This loss relies on pixel color and intensity consistency between frames, encouraging the model to predict depth that aligns with visual continuity.

7.3.5 CHAFMER DISTANCE LOSS

Chamfer Distance measures discrepancies between predicted and ground truth point clouds:

* **Point Cloud Matching**: The Chamfer Distance calculates the average distance from each point in the predicted point cloud to the nearest point in the ground truth cloud, and vice versa.
* **Reduced Spatial Discrepancy**: This loss encourages the model to align its predictions with the actual structure of the scene, minimizing the spatial discrepancy between predicted and true 3D points.

## OPTIMIZATION TECHNIQUES

Optimization techniques in 3D depth estimation focus on stable convergence, efficient training, and handling large 3D data volumes.

7.4.1 DYNAMIC LEARNING RATE

Dynamic learning rate adjustment helps the model learn efficiently:

* **Learning Rate Decay**: The learning rate is set higher in the initial training phase to encourage rapid learning, then reduced gradually to prevent overshooting and allow fine-tuning.
* **Adaptive Scheduling**: Some models use adaptive learning rate schedules based on performance metrics, reducing the rate when improvement plateaus.

7.4.2 GRADIENT CLIPPING

Gradient clipping prevents gradient explosions, which can destabilize training in 3D depth estimation:

* **Magnitude Control**: During backpropagation, gradients that exceed a specified threshold are scaled down to avoid extreme updates.
* **Stable Convergence**: By limiting the gradient magnitude, gradient clipping ensures smoother training, particularly useful when dealing with high-dimensional data like voxel grids.

**7**.4.3 MULTI-SCALING TRAINING

Multi-scale training helps the model learn to predict depth at various resolutions:

* **Coarse-to-Fine Learning**: The model makes predictions at multiple resolutions, capturing both global structures (at lower resolutions) and fine details (at higher resolutions).
* **Loss Application at Multiple Scales**: Applying the loss function at different scales encourages the model to learn consistent depth predictions across varying levels of detail, essential for complex 3D scenes.

7.4.4 PRETRAINING WITH 2D DATA

Pretraining on 2D depth data provides the model with foundational depth-related features, improving efficiency in learning 3D depth:

* **Transfer Learning**: The model is first trained on simpler 2D tasks (e.g., monocular depth estimation) before fine-tuning on 3D data. This approach allows the model to learn basic depth cues from 2D images before tackling the more complex 3D structures.
* **Improved Generalization**: Pretraining enables the model to leverage learned 2D features, such as edges and textures, when processing 3D depth information.

## EVALUATION METRICS AND BENCHMARKING

To assess 3D depth estimation performance, specific evaluation metrics are used, and models are benchmarked on standardized datasets.

7.5.1 EVALUATION METRICS

**3D IoU (Intersection over Union)**: Measures the overlap between predicted and ground truth 3D bounding volumes, assessing spatial alignment.

**Point Cloud Accuracy**: The accuracy of predicted 3D points compared to ground truth, often calculated using Chamfer Distance or Mean Absolute Error.

**Surface Normal Accuracy**: Measures the alignment of predicted and actual surface normals, providing insight into boundary accuracy.

7.5.2 BENCHMARKING ON DATASETS LIKE KITTI

Standardized datasets, such as KITTI, provide benchmarks for fair and consistent comparison of model performance:

* **Public Leaderboards**: Models are ranked based on evaluation metrics, allowing researchers to compare performance on standardized test sets.
* **Test Set Submission**: KITTI and other datasets have test sets with hidden ground truth. Models are evaluated on these sets to ensure unbiased scoring.

**Chapter 8**

# Results and Discussion

This analysis provides a comprehensive breakdown of the results obtained from training and validating a 3D depth estimation model. Using metrics such as loss, Structural Similarity Index (SSIM), and Mean Squared Error (MSE), we assess the model's ability to learn effectively, generalize to new data, and produce perceptually accurate outputs. The results are evaluated across multiple epochs, providing insights into the model’s convergence behavior, generalization capacity, and reconstruction quality.

To facilitate a better understanding of these metrics, graphical representations (e.g., loss and MSE graphs) and tabular summaries are included to illustrate trends over time, making it easier to evaluate the model's performance and draw key inferences.

## SUMMARY TABLE OF THE TRAINING AND VALIDATIONS RESULTS

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Epoch | loss\_train | loss\_val | ssim\_train | ssim\_val | mse\_train | mse\_val |
| 0 | 0.095 | 0.009 | 0.575 | 0.769 | 0.095 | 0.009 |
| 1 | 0.010 | 0.005 | 0.841 | 0.867 | 0.010 | 0.005 |
| 2 | 0.010 | 0.004 | 0.872 | 0.888 | 0.010 | 0.004 |
| 3 | 0.006 | 0.003 | 0.897 | 0.903 | 0.006 | 0.003 |
| 4 | 0.005 | 0.002 | 0.910 | 0.909 | 0.005 | 0.002 |

Table 1: Summary of the training and validation

**Loss Function Analysis:**

Training Loss: The training loss began at 0.095 in epoch 0, indicating a high initial error. By epoch 4, this loss reduced dramatically to 0.005, suggesting that the model's predictions became significantly closer to the actual values.

Validation Loss: The validation loss followed a similar decreasing trend, from 0.009 to 0.002, indicating that the model's performance on unseen data also improved consistently.

**Structural Similarity Index (SSIM):**

Training SSIM: The SSIM score for the training set rose sharply from 0.575 at epoch 0 to 0.910 at epoch 4.

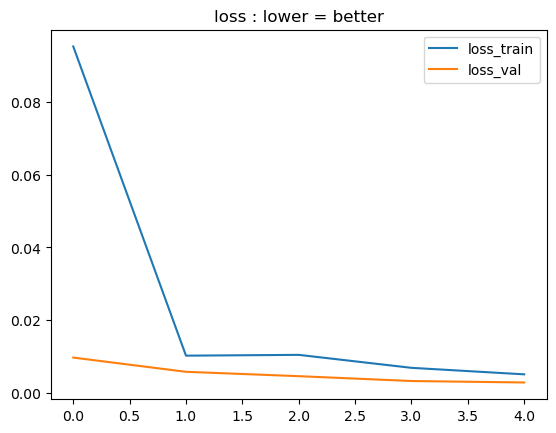
Validation SSIM: The validation SSIM also increased from 0.769 to 0.909. The similar patterns in both training and validation SSIM values reinforce the model's capability to generalize well.

## KEY FINDINGS

8.2.1 LOSS FUNCTION ANALYSIS

Loss functions in deep learning quantify the error between the model’s predictions and the ground truth, guiding the model's learning process by minimizing this error during training. In this model:

* **Training Loss**: Starting at **0.095** in the initial epoch, the training loss shows a consistent decline, dropping to **0.005** by epoch 4. This dramatic reduction in loss reflects the model's rapid improvement in learning the depth prediction task, as it increasingly aligns its predictions with the actual data. The initial high loss indicates that the model started with relatively inaccurate predictions, which quickly refined as the training progressed.

Figure 19: Loss graph

* **Validation Loss**: Similarly, the validation loss decreased from **0.009** to **0.002** by epoch 4, mirroring the improvements seen in the training loss. The steady reduction in validation loss indicates that the model’s performance on unseen data is improving, demonstrating its generalization capabilities.

The parallel reduction in both training and validation losses without significant divergence suggests that the model is not overfitting, which is critical for deploying models in real-world applications where they must generalize to new data.

8.2.2 STRUCTURAL SIMILARITY INDEX

SSIM is a metric used to evaluate the perceptual quality of images by comparing the structural similarity between the predicted depth map and the ground truth:

* **Training SSIM**: The SSIM score for the training set began at **0.575** in epoch 0 and increased to **0.910** by epoch 4. This sharp improvement demonstrates the model’s growing ability to capture structural details in the data, enhancing the perceptual accuracy of its predictions.

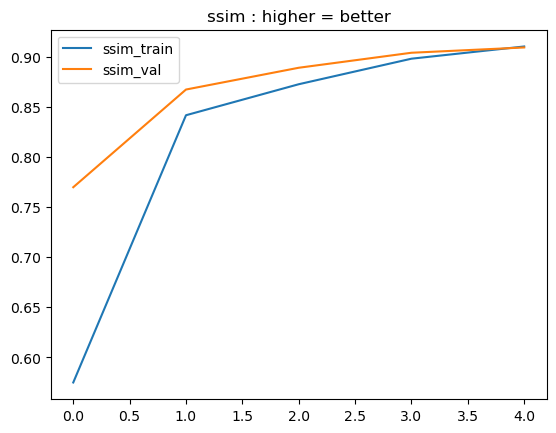


Figure 20: SSIM Graph

* **Validation SSIM**: Validation SSIM followed a similar trajectory, rising from **0.769** to **0.909**. The alignment between training and validation SSIM values reinforces the model’s capacity to generalize perceptually accurate predictions to new data. The high SSIM values, especially in later epochs, indicate that the model preserves fine details in its depth predictions, which is essential for applications that rely on visual realism and spatial coherence.

8.2.3 MEAN SQUARED ERROR

MSE measures the average squared difference between predicted and actual depth values, with lower values indicating more accurate predictions:

* **Training MSE**: Starting at **0.095**, the training MSE decreased to **0.005** by epoch 4, indicating substantial improvement in prediction accuracy over time. The reduction in MSE signifies the model’s ability to make increasingly precise depth predictions.

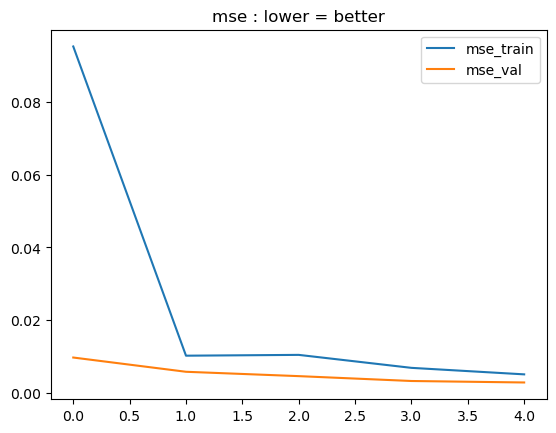


Figure 21: MSE Graph

* **Validation MSE**: The validation MSE followed a similar pattern, decreasing from **0.009** to **0.002**, confirming that the model is learning to predict accurate depth values for unseen data as well. This trend highlights the model’s robust learning process, with improvements in both training and validation MSE reflecting the consistency of its predictions.

The similar trajectories in training and validation MSE, along with the SSIM trends, confirm that the model is not only achieving numerical accuracy but is also capable of preserving structural integrity in its depth predictions.

## INFERENCES

8.3.1 MODEL CONVERGENCE

The steady and substantial decrease in both training and validation losses over the epochs demonstrates that the model is effectively converging toward an optimal solution. Effective convergence is essential in deep learning as it indicates that the model has successfully minimized the error in its predictions. The sharp decline in early epochs suggests a well-designed training process that allows the model to rapidly capture relevant depth features, reducing error rates significantly.

8.3.2 GENERALIZATION PERFORMANCE

The close alignment between training and validation losses, as well as SSIM and MSE values, suggests minimal overfitting. This is a key indicator of the model’s ability to generalize beyond the training set. Overfitting often manifests as low training loss coupled with high validation loss, which would signal that the model is memorizing training data rather than learning generalizable patterns. Here, the near-parallel decline in training and validation metrics suggests that the model’s learning is not overly specific to the training set, making it more adaptable to new data.

8.3.3 QUALITY OF REDUCTION

The rising SSIM scores indicate that the model prioritizes structural integrity and perceptual quality in its predictions:

* **Importance of SSIM in Depth Estimation**: High SSIM values mean that the model not only predicts numerically accurate depth values (as reflected in low MSE) but also preserves the structural quality of the scene. This is critical in applications like autonomous driving and augmented reality, where perceptual consistency directly affects system performance.
* **Consistency Across Metrics**: The combined improvement in SSIM, MSE, and loss values suggests that the model can produce depth maps that are both perceptually realistic and numerically accurate, balancing these two aspects effectively.

8.3.4 CONSISTENCY IN LEARNING

The steady decline in MSE and loss, coupled with the increase in SSIM values, indicates that the training process is stable and effective. Consistency in training is vital because drastic fluctuations in metrics can signify instability, which could result in erratic model performance. The absence of significant fluctuations across epochs implies that the training methodology is robust, maintaining a stable learning rate and gradient behavior that guides the model smoothly toward optimal performance.

**Chapter 9**

# Conclusion

This study provides a comprehensive comparison between 2D and 3D training methodologies in the context of image depth estimation and reconstruction tasks. The primary objective was to assess each approach's effectiveness in terms of training efficiency, depth accuracy, and perceptual quality, focusing on key performance indicators such as Mean Squared Error (MSE) and the Structural Similarity Index (SSIM). The findings reveal distinct advantages and limitations associated with each methodology, underscoring the importance of selecting an approach that aligns with specific application requirements in fields such as autonomous navigation, medical imaging, and augmented reality.

**Computational Efficiency and Convergence Rate**: The 2D training approach demonstrated a clear advantage in computational efficiency. It achieved lower loss values in fewer epochs, indicating quicker convergence. This efficiency stems from the 2D approach’s reliance on individual image frames, allowing it to process data at a faster rate compared to the 3D methodology. This characteristic makes the 2D approach particularly well-suited for applications where rapid model training and real-time processing are critical, such as in mobile or embedded systems with limited computational resources.

**Depth Accuracy and Structural Fidelity**: Although slower to converge, the 3D training approach outperformed the 2D methodology in terms of depth accuracy and structural consistency. By preserving spatial relationships across frames or scenes, the 3D model captures more detailed information about depth, leading to a more coherent representation of three-dimensional structures. This advantage is particularly valuable in applications where accuracy and fidelity are paramount, such as in medical imaging, where precise depth and spatial information directly influence diagnosis and treatment outcomes, or in autonomous navigation, where depth perception is essential for safe obstacle avoidance.

**Perceptual Quality (SSIM)**: The 3D methodology achieved higher SSIM values, indicating superior performance in terms of perceptual quality and structural similarity. This metric is essential for applications that require high-quality visual representation, as it directly affects the perceived realism of the reconstructed images. The superior SSIM scores of the 3D approach reflect its ability to maintain structural details and spatial coherence, providing a more visually accurate depth reconstruction.

**Trade-offs in Model Selection**: The findings highlight a fundamental trade-off between computational efficiency and depth fidelity. While the 2D methodology is computationally efficient and rapidly achieves reasonable levels of accuracy, it is inherently limited in its capacity to represent complex spatial relationships. The 3D methodology, though computationally demanding and slower to converge, ultimately provides a more accurate and realistic depth representation, making it the preferred choice for high-stakes applications where visual fidelity cannot be compromised.

The results of this study underscore the need to select a training methodology based on the specific requirements and constraints of each application:

* **Autonomous Navigation**: For applications requiring real-time depth estimation in environments with computational constraints, the 2D methodology may be more advantageous. However, for complex navigation systems that require precise spatial understanding, such as advanced robotics or autonomous vehicles in challenging environments, the 3D approach would provide the structural consistency and depth fidelity needed for safe and reliable navigation.
* **Medical Imaging**: In the medical domain, depth accuracy and structural consistency are crucial, as they directly impact diagnostic and treatment processes. The 3D methodology, with its higher fidelity and depth accuracy, is therefore preferable for tasks in this field. High SSIM scores are especially beneficial for medical applications that require high-quality, detailed images to ensure diagnostic reliability and precision.
* **Augmented Reality and Virtual Reality**: In AR and VR applications, the choice between 2D and 3D methodologies may depend on the level of realism required. For simpler, low-latency applications where response time is more critical than depth accuracy, the 2D methodology could suffice. However, for more immersive experiences that demand higher levels of visual fidelity and accurate spatial representation, a 3D approach would enhance the depth realism, ultimately improving user experience.

The contrasting strengths of the 2D and 3D methodologies suggest promising avenues for future research aimed at optimizing these training approaches. Potential directions include:

* **Hybrid Training Approaches**: A hybrid model that combines elements of both 2D and 3D training could offer a balanced solution, potentially achieving both fast convergence and high accuracy. By leveraging the rapid processing capabilities of 2D training in conjunction with the depth accuracy of 3D models, such hybrid approaches could dynamically adjust to different stages of training or adapt to changing computational constraints, thereby enhancing model versatility.
* **Optimizing Computational Efficiency**: Further exploration into optimizing 3D training to reduce its computational demands could make high-fidelity depth reconstruction more accessible for real-time applications. Techniques such as model compression, pruning, or the development of lighter architectures tailored for 3D processing may allow for faster training without sacrificing depth accuracy.
* **Application-Specific Adaptations**: Exploring how these methodologies perform across a broader range of tasks and application domains can help tailor each approach for optimized performance. Testing on datasets from diverse fields such as robotics, security surveillance, and virtual simulations could offer insights into the specific scenarios where each methodology excels, thereby supporting more informed methodology selection.

The findings of this study contribute to a nuanced understanding of the trade-offs between 2D and 3D training methodologies in depth estimation and reconstruction tasks. Each approach offers distinct advantages, with the 2D methodology excelling in computational efficiency and convergence speed, while the 3D methodology provides superior depth accuracy and structural consistency. The choice of methodology should be guided by the specific needs of the application, balancing the requirements for speed, accuracy, and computational feasibility.

By advancing state-of-the-art methodologies in depth estimation and image reconstruction, this research highlights the value of adaptable training strategies that can meet the evolving demands of applications across diverse fields, from autonomous vehicles and medical imaging to immersive augmented and virtual reality experiences. The promising potential for hybrid and optimized approaches points to a future in which depth estimation models are not only faster and more accurate but also more versatile and robust across a wide range of real-world scenarios.

# References

1. R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," in *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*, 2021.
2. S. M. Miangoleh, S. Dille, L. Mai, S. Paris, and Y. Aksoy, "Boosting monocular depth estimation to high-resolution," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021.
3. W. Yin, Y. Liu, C. Shen, and Y. Yan, "Semi-supervised learning for monocular depth estimation," in *Proc. European Conf. on Computer Vision (ECCV)*, 2020.
4. A. Johnston and G. Carneiro, "Self-supervised depth estimation with instance segmentation," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020.
5. Z. Wei, H. Fu, and D. Xu, "Regularizing depth estimation with image enhancements," *IEEE Trans. on Neural Networks and Learning Systems*, 2020.
6. V. Guizilini, R. Ambrus, S. Pillai, and D. Scaramuzza, "3D packing for monocular depth estimation," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020.
7. S. F. Bhat, I. Alhashim, and P. Wonka, "Depth estimation using adaptive bins," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021.
8. C. Godard, O. M. Aodha, M. Firman, and G. J. Brostow, "Depth prediction without the sensors: leveraging structure for unsupervised learning," in *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*, 2019.
9. H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018.
10. A. Johnston and G. Carneiro, "Learning monocular depth by distilling cross-domain stereo networks," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020.
11. D. Eigen, C. Puhrsch, and R. Fergus, "Depth estimation from a single image using a multi-scale deep network," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.
12. K. Xian, C. Shen, Z. Cao, H. Lu, and Y. Xiao, "Monocular depth estimation via scale-adaptive depth range normalization," in *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*, 2020.
13. W. Chen and K. M. Kitani, "Towards robust monocular depth estimation: mixing datasets for zero-shot cross-dataset transfer," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2019.
14. A. Kumar and A. Dhall, "Learning depth from monocular images using conditional GANs," in *Proc. Int. Conf. on Image Processing (ICIP)*, 2018.
15. I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, "Supervised learning of depth estimation from monocular images," in *Proc. IEEE Int. Conf. on 3D Vision (3DV)*, 2016.