



אוניברסיטת בן-גוריון בנגב

הפקולטה למדעי הטבע

המחלקה להנדסת מכונות

## **בחינת אלגוריתמים לשערוך עומק באמצעות מצלמות סטריאו**

דו"ח סופי

מספר פרויקט : 24-123

רז תורג'מן - 315047316

בהנחיית ד"ר אור צליל

חתימה : \_\_\_\_\_



Ben-Gurion University of the Negev  
The Faculty of Natural Sciences  
The Department of **Mechanical Engineering**

# Examining algorithms for depth estimation using stereo cameras

Final report  
Project number: 24-123

**Raz Turgeman - 315047316**

Under the supervision of **Dr. Or Tslil**

Signature: \_\_\_\_\_

September 5, 2024

2 Elul 5784

# Supervisor's Opinion

The final project titled "*Examining Algorithms for Depth Estimation Using Stereo Cameras*" by Raz Turgeman at Ben-Gurion University of the Negev tackles a significant challenge in computer vision: the development and evaluation of depth estimation methods using stereo camera setups. In this project, Raz has explored a range of depth estimation techniques, including both traditional stereo matching and advanced neural network models, such as HITNET and CRE, to assess their effectiveness in various environments.

Raz implemented these models using the Intel RealSense D455 stereo camera, leveraging its hardware-based stereo matching and integrating neural network-based depth estimation models to enhance performance. The project included rigorous frame-by-frame and pixel-wise error analysis, using triangulation as a ground truth method to compare and validate the depth estimates generated by each model.

The results of this project highlighted the strengths of neural network-based models over traditional methods, particularly in complex environments involving dynamic lighting and varying textures. Raz's evaluation showed that the HITNET and CRE models significantly reduced errors and improved depth map accuracy, outperforming the Intel D455's built-in stereo matching in various test scenarios.

Throughout the project, Raz demonstrated outstanding technical skills and a deep understanding of stereo vision and depth estimation. His approach to implementing and comparing these algorithms was methodical and thorough, leading to meaningful insights into the performance of each model under different conditions. His work in integrating both traditional and modern depth estimation methods sets a solid foundation for further research in this field.

Raz's dedication to the project, his methodical problem-solving abilities, and his capacity for independent research are evident in the quality of his work. His contribution to the field of depth estimation using stereo cameras is significant, and his project has the potential for real-world applications in areas such as robotics, autonomous navigation, and

industrial automation. Raz's project reflects his strong technical skills and his potential for future success in engineering and computer vision.

**Dr. Or Tzlil**

**Elbit Systems**

# תקציר

תחום הראייה הממוחשבת ועיבוד התמונה התקדם בשנים האחרונות הן במחקר והן בפיתוח. אתגר מרכזי בתחום זה הוא שערור עומק באמצעות זוג מצלמות סטריאו בזמן אמת תוך שאיפה לאיכות תמונה ואמינות גבוהה. שערור עומק בזמן אמת הינו תהליך של קביעת מרחק של אובייקטים במרחב מנקודת המבט של המצלמות מתוך המישור הדו ממדי של המצלמות הנתון בפיקסלים.

אחד האתגרים בשערור עומק באמצעות מצלמות סטריאו הוא ביצוע קליברציה למצלמות שבעזרתה יופקו פרמטרים הקושרים בין עדשת המצלמה ומישור התמונה - שלה אינטרינזיקה, ופרמטרים הקושרים בין המצלמה לנקודה יחוס בעולם - אקסטרנזיקה. באמצעות פרמטרים אלו ניתן להתאים אובייקטים בין שתי התמונות המתקבלות מהמצלמות וליצור map disparity שבאמצעותה ניתן לשערך את העומק. בזכות השימוש במצלמת העומק D455 Realsense Instel שבה נעשה שימוש בפרויקט, הדיוק בין עדשות המצלמה מספק והיא מגיעה עם אמצעים לכיול עצמי מה שהופך את התהליך למהיר ונוח וחוסך בשגיאות ביחס לייצור עצמי כמו למשל בהדספת תלת מימד. אתגר נוסף ועיקרי הינה קבלת תוצאה שאינה רועשת, ישנו מגוון רחב של אלגוריתמים בהם ניתן להשתמש כדי להפיק את המפה הדרושה לשערור עומק, אשר חלקם מפיקים תוצאה רועשת וזו תוצאה שיש למזער ככל הניתן.

פרויקט זה מציע מימוש ובחינה של אלגוריתמים לשערור עומק באמצעות מצלמות סטריאו, תוך התמודדות עם האתגרים המשמעותיים הקשורים לתמונות רועשות. מצלמות תרמיות, למרות שהן מועילות ליישומים שונים, לעיתים קרובות מפיקות תמונות עם רעש משמעותי, מה שמסבך את המשימה של הערכת עומק. כדי להתגבר על בעיה זו, נרצה לממש שיטות מתקדמות כמו רשתות נוירונים המפיקות רמות עומק ברמה גבוהה ועל מנת לקבוע את איכותן יבוצע ניתוח של התוצאות ביחס לפיקסלים עבורם נוכל לומר בביטחון גבוהה את העומק המקורב של הנקודה. משפט הקיום של הפרויקט הינו מימוש אלגוריתם שערור עומק באמצעות מצלמות סטריאו בזמן אמת ויישום כלים של בינה מלאכותית כפתרון לבעית התמונה הרועשת תוך הערכה של איכותן.

המסקנה העיקרית של הפרויקט הינה שעל מנת שתתקבל תמונה ברורה איכותית וללא רעשים ככל הניתן, האלגוריתמים הפשוטים אינם מספקים, ויש להשתמש בשיטות וכלים מתקדמים שמשלבים רשתות נוירונים על מנת שנוכל לקבל תוצאה איכותית ומהימנה. בנוסף, כלים להערכת איכות השערור מהרשתות הם חשובים על מנת לקבוע אם השימוש ברשת מסוימת אכן מספק את צרכי המוצר עליו מיישמים את השימוש במפות העומק. איכות הרשמת מושפעת מפונקציית המחיר שלה, סט המידע עליו היא אומנ ומכאן אופן ההתמודדות עם תנאים משתנים כמו סביבת פנים וחוף, תאורה ועוד.

# Abstract

The field of computer vision and image processing has seen significant advancements in both research and development in recent years. A central challenge in this domain is real-time depth estimation using stereo camera pairs, with an emphasis on achieving high image quality and reliability.

Real-time depth estimation involves determining the distance of objects in space from the perspective of the cameras, based on the two-dimensional image plane captured in pixels. One of the primary challenges in depth estimation using stereo cameras is performing calibration. This calibration process yields intrinsic parameters that link the camera lens to its image plane and extrinsic parameters that relate the camera to a reference point in the world. By using these parameters, it is possible to match objects between the two images obtained from the cameras and generate a disparity map, which is then used to estimate depth. The Intel RealSense D455 depth camera, utilized in this project, offers precise lens alignment and includes self-calibration features, making the process faster and reducing errors compared to custom setups, such as those involving 3D-printed components.

Another major challenge is obtaining noise-free results. There is a wide range of algorithms available for generating the necessary maps for depth estimation, but some produce noisy outputs that must be minimized as much as possible.

This project proposes the implementation and evaluation of algorithms for depth estimation using stereo cameras, addressing the significant challenges associated with noisy images. While thermal cameras are beneficial for various applications, they often produce images with significant noise, complicating the task of depth estimation. To overcome this challenge, advanced methods such as neural networks are employed to generate high-quality depth maps. The quality of these depth maps is then analyzed relative to the pixels where the approximate depth is known with high confidence. The central aim of the project is the real-time implementation of a depth estimation algorithm using stereo

cameras and the application of neural networks models as a solution to the problem of noisy images, alongside an assessment of their quality.

The main conclusion of this project is that simple algorithms are insufficient for producing clear, high-quality, and noise-free images. Advanced methods and tools that incorporate neural networks are necessary to achieve reliable and high-quality results. Additionally, tools for assessing the quality of depth estimation from these networks are crucial to determine whether a particular network meets the requirements of the product in which the depth maps are used. The quality of the estimation is influenced by the network's cost function, the dataset on which it is trained, and its ability to handle varying conditions such as indoor and outdoor environments, lighting, and more.

Overall, the CRE model provided the best results across the various scenarios analyzed, demonstrating superior accuracy and reliability in depth estimation compared to other methods.

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Overview and Motivation . . . . .	1
1.2	Project Goals . . . . .	1
1.3	Project Scope . . . . .	2
1.4	Importance of the Project . . . . .	3
1.5	Structure of the Work and Main Contributions . . . . .	3
<b>2</b>	<b>Camera Calibration</b>	<b>5</b>
2.1	Intel RealSense D455 Calibration . . . . .	5
<b>3</b>	<b>RealSense D455 Depth Estimation</b>	<b>6</b>
3.1	Filtering Techniques . . . . .	6
3.2	Infrared Emitter . . . . .	7
<b>4</b>	<b>Triangulation</b>	<b>8</b>
4.1	Algorithm Development . . . . .	8
4.2	Mathematical Foundations . . . . .	8
4.3	Application as Sparse Ground Truth . . . . .	9
<b>5</b>	<b>Neural Network Depth Estimation</b>	<b>11</b>
5.1	Model Implementation . . . . .	11
5.2	CRE and HITNET Models . . . . .	11
5.2.1	CRE Model . . . . .	12
5.2.2	HITNET Model . . . . .	12



5.3	Comparison of CRE and HITNET . . . . .	14
<b>6</b>	<b>Depth Evaluation Process</b>	<b>15</b>
6.1	Evaluation Methods . . . . .	15
6.2	Mathematical Formulation of Metrics . . . . .	16
6.2.1	Mean Absolute Error (MAE) . . . . .	16
6.2.2	Root Mean Squared Error (RMSE) . . . . .	16
6.2.3	Mean Squared Error (MSE) . . . . .	16
6.2.4	Intersection over Union (IoU) . . . . .	17
6.2.5	Percentage of Erroneous Pixels (D1) . . . . .	17
6.3	Output Graphs and Comparisons . . . . .	17
6.4	Critical Analysis and Conclusions . . . . .	18
<b>7</b>	<b>Results Analysis</b>	<b>19</b>
7.1	Outdoor Recording . . . . .	19
7.1.1	Frame-by-Frame Analysis . . . . .	19
7.1.2	Overall Outdoor Analysis . . . . .	24
7.2	Indoor Recording . . . . .	27
7.2.1	Frame-by-Frame Analysis . . . . .	27
7.2.2	Overall Indoor Analysis . . . . .	32
7.3	Comparative Analysis . . . . .	32
7.3.1	Comparison between Indoor and Outdoor Recordings . . . . .	32
7.4	Conclusion . . . . .	34
<b>8</b>	<b>Project Expenses</b>	<b>37</b>
<b>9</b>	<b>Gantt Chart</b>	<b>38</b>

## A Appendix

A.1	Project Repository . . . . .	
A.2	Indoor Data Analysis . . . . .	
A.2.1	Error Progression Over Time . . . . .	
A.3	Outdoor Data Analysis . . . . .	
A.3.1	Error Progression Over Time . . . . .	

# List of Figures

4.1	Example of triangulation algorithm, showing the sparse depth measurements derived from stereo images. . . . .	10
5.1	Example of CRE (a) and HITNET (b), showing the dense estimated depth.	14
7.1	Depth comparison for frame 35 in outdoor recording. . . . .	19
7.2	Difference between HITNET, CRE, and D455 for frame 35 in outdoor recording. . . . .	20
7.3	Error distribution for frame 35 in outdoor recording. . . . .	21
7.4	Depth comparison for frame 156 in outdoor recording. . . . .	22
7.5	Difference between HITNET, CRE, and D455 for frame 156 in outdoor recording. . . . .	23
7.6	Error distribution for frame 156 in outdoor recording. . . . .	24
7.7	Depth comparison for frame 97 in indoor recording. . . . .	27
7.8	Difference between HITNET, CRE, and D455 for frame 97 in indoor recording. . . . .	28
7.9	Error distribution for frame 97 in indoor recording. . . . .	29
7.10	Depth comparison for frame 179 in indoor recording. . . . .	30
7.11	Difference between HITNET, CRE, and D455 for frame 179 in indoor recording. . . . .	30
7.12	Error distribution for frame 179 in indoor recording. . . . .	31
A.1	D1 Error Progression Over Time for Indoor Data . . . . .	
A.2	IoU Error Progression Over Time for Indoor Data . . . . .	
A.3	MAE Error Progression Over Time for Indoor Data . . . . .	

A.4	MSE Error Progression Over Time for Indoor Data . . . . .
A.5	RMSE Error Progression Over Time for Indoor Data . . . . .
A.6	Error Comparison for All Frames for Indoor Data . . . . .
A.7	D1 Error Progression Over Time for Outdoor Data . . . . .
A.8	IoU Error Progression Over Time for Outdoor Data . . . . .
A.9	MAE Error Progression Over Time for Outdoor Data . . . . .
A.10	MSE Error Progression Over Time for Outdoor Data . . . . .
A.11	RMSE Error Progression Over Time for Outdoor Data . . . . .
A.12	Error Comparison for All Frames for Outdoor Data . . . . .

# List of Tables

7.1	Metrics Summary for Outdoor Recording . . . . .	25
7.2	Metrics Summary for Indoor Recording . . . . .	32
8.1	Project Expenses . . . . .	37

# 1 Introduction

## 1.1 Overview and Motivation

The field of computer vision has experienced significant advancements, particularly in image processing and depth estimation [15]. Depth estimation is crucial for various applications, including autonomous vehicles, robotics, augmented reality, and industrial automation [13]. By accurately measuring the distance to objects in a scene, these systems can interact with their environment more effectively and safely. Stereo vision, which involves using two cameras to capture slightly different perspectives of the same scene, is a widely used technique for depth estimation [12]. Analyzing the differences (disparity) between the two images allows for inference of depth information. However, accurate depth estimation with stereo cameras is challenging, especially when state-of-the-art results are required.

## 1.2 Project Goals

The primary goal of this project is to implement and evaluate algorithms for depth estimation using stereo cameras.

Specific goals include:

- Develop a real-time depth estimation algorithm using stereo cameras.
- Implement camera calibration techniques to obtain accurate intrinsic and extrinsic parameters [2].
- Explore advanced methods and tools, including neural networks and artificial intelligence, to enhance depth estimation accuracy [17].
- Evaluate the performance of different depth map generation algorithms.

## 1.3 Project Scope

This project encompasses several key phases, each focused on advancing the accuracy and reliability of depth estimation using the Intel RealSense D455 stereo depth camera and advanced neural network algorithms:

- **Camera Setup and Calibration:** The Intel RealSense D455 stereo depth camera was selected for its precision and industrial-grade build quality, ensuring that all lenses are perfectly aligned during the manufacturing process. This camera comes equipped with on-chip calibration, which provides intrinsic and extrinsic parameters automatically, eliminating the need for external calibration tools [2]. This feature streamlines the setup process and guarantees consistent and reliable depth data from the outset.
- **Stereo Matching Depth Estimation:** Leveraging the Intel RealSense D455's advanced built-in stereo matching algorithm, the project bypasses the need for traditional disparity map generation methods. The camera's integrated technology provides real-time depth estimation with filtering and laser emitter to reduce noise, enabling the focus to shift toward analyzing and enhancing the output quality rather than implementing basic stereo matching algorithms [2].
- **Triangulation Algorithm:** The project includes the development of a triangulation algorithm that calculates depth by leveraging the rectified stereo images from the Intel RealSense D455 camera [5]. This algorithm uses the known baseline distance between the camera lenses and the disparity between corresponding pixels in the left and right images to compute depth, providing a sparse ground truth comparison for the neural network models and the camera's built-in depth estimation.
- **Neural Network Depth Estimation:** The project explores state-of-the-art neural network models like HITNET and CRE to further optimize depth estimation [17]. These models are implemented and tested on data obtained from the Intel RealSense D455, with the goal of comparing their performance against the camera's built-in methods.

- **Evaluation and Optimization:** The performance of both traditional and neural network-based depth estimation algorithms is evaluated using a variety of metrics, including accuracy, reliability, and robustness to noise [10]. The project includes an in-depth analysis of the results obtained from both indoor and outdoor environments, providing insights into the strengths and limitations of each approach. The CRE model, in particular, has been identified as providing the best overall results, demonstrating superior accuracy and reliability across various conditions [17].

## 1.4 Importance of the Project

This project tackles a crucial challenge in computer vision: achieving accurate and reliable depth estimation using stereo cameras in a variety of environmental conditions [15]. By developing and evaluating advanced depth estimation algorithms, including neural network models, this work contributes to the improvement of stereo vision systems. The results of this project have the potential to enhance the accuracy and reliability of depth estimation across a wide range of industries and applications, leading to better performance in tasks requiring precise environmental interaction [6].

## 1.5 Structure of the Work and Main Contributions

The structure of the work carried out in this project is as follows:

- **Phase 1: Camera Setup and Calibration:** Selection and setup of the Intel RealSense D455 stereo camera, utilizing its on-chip calibration capabilities to obtain precise intrinsic and extrinsic parameters [2]. This phase also included designing and 3D printing a custom camera platform to ensure stable and accurate image capture.
- **Phase 2: Ground Truth Algorithm Development:** A triangulation algorithm was developed to calculate depth using rectified stereo images from the D455 camera [5]. This algorithm provided a sparse ground truth for evaluating the accuracy of both the neural network models and the camera’s built-in depth estimation methods, serving as a critical reference point throughout the project.



- **Phase 3: Neural Network Depth Estimation:** The project implemented and evaluated state-of-the-art neural network models, including HITNET and CRE, specifically for depth estimation [17]. These models were tested on data collected from the D455 camera, with the goal of comparing their performance against the camera’s built-in stereo matching algorithm, ultimately seeking to determine the most effective method for accurate depth estimation.
- **Phase 4: Depth Evaluation Algorithm and Data Analysis:** A comprehensive evaluation was conducted to assess the performance of the various depth estimation methods [10]. Metrics such as accuracy, processing time, and robustness to noise were used to analyze the results. The analysis led to the identification of the CRE model as the best overall performer, highlighting its superior accuracy and reliability across different testing conditions [17].

The main contributions of this project include the development of a reliable triangulation algorithm, the application and assessment of neural networks for enhanced depth accuracy, and an in-depth analysis of depth estimation methods across different environments. This work provides a solid foundation for further advancements in depth estimation technologies, with implications for a variety of real-world applications.

## 2 Camera Calibration

Camera calibration is a critical step in the process of depth estimation, as it ensures the accuracy of the 3D reconstruction of the scene [18]. The primary goal of camera calibration is to determine the camera’s intrinsic and extrinsic parameters [8]. Intrinsic parameters include the focal length, principal point, and distortion coefficients, which define how the camera interprets a 3D scene into a 2D image [18]. Extrinsic parameters include the rotation and translation vectors, which describe the camera’s position and orientation in space [1].

This calibration process is crucial for depth estimation because accurate camera parameters allow for the correct triangulation of points in 3D space. Miscalibration can lead to significant errors in depth measurement, making it essential to ensure that the calibration is both precise and accurate [4].

### 2.1 Intel RealSense D455 Calibration

The Intel RealSense D455 camera simplifies the calibration process by integrating an on-chip calibration mechanism [3]. This process is carried out internally by the camera and is designed to automatically adjust the camera’s intrinsic and extrinsic parameters to ensure optimal depth accuracy. The on-chip calibration method employed by the D455 involves the camera analyzing specific internal reference patterns and adjusting its parameters accordingly, without requiring user intervention [3].

## 3 RealSense D455 Depth Estimation

Stereo matching is a fundamental technique in computer vision for estimating depth from a pair of stereo images. The principle behind stereo matching is to find corresponding points in the left and right images that represent the same point in the 3D scene [12]. Once these correspondences are found, the disparity, or the difference in pixel locations between the two images, can be calculated. This disparity is inversely proportional to the depth of the object in the scene, allowing for the reconstruction of the 3D structure.

The camera utilizes a stereo matching algorithm integrated within its hardware to perform real-time depth estimation. Leveraging its global stereo matching technique, the D455 efficiently processes stereo images to produce dense depth maps with high accuracy. The accuracy and reliability of the depth estimation are further enhanced through the application of filters, as well as the use of the camera's infrared (IR) emitter [3].

### 3.1 Filtering Techniques

The following filters are enabled and configured to optimize the depth data obtained from the D455 [2]:

- **Decimation Filter:** This filter reduces the resolution of the depth map by down-sampling the image. The decimation process helps remove redundant data, thus reducing noise and enhancing the processing speed. By focusing on the most critical data points, this filter improves the clarity and accuracy of the depth map without significantly sacrificing detail [3].
- **Spatial Filter:** The spatial filter works by smoothing the depth map, taking into account the depth values of neighboring pixels. This filter applies a weighted average to these values, reducing high-frequency noise and producing a more uniform depth map [9]. The spatial filter is particularly useful in eliminating small, isolated artifacts that can arise in the depth data, resulting in a cleaner, more consistent depth estimation [2].

- **HDR Merge:** The High Dynamic Range (HDR) merge function is enabled to combine depth data captured at different exposure levels. This process enhances the depth map’s detail and accuracy by balancing the information obtained from both low and high-exposure images. It is particularly effective in scenes with varying lighting conditions, ensuring that both bright and dark areas are accurately represented in the final depth map [2].

## 3.2 Infrared Emitter

The D455’s IR emitter plays a crucial role in improving the performance of the stereo matching algorithm, especially in challenging lighting conditions [2]. The emitter projects a structured light pattern onto the scene, which enhances the texture on otherwise textureless surfaces. This additional texture allows the stereo matching algorithm to identify corresponding points more effectively, even on surfaces that would typically be difficult to match due to low contrast or lack of detail [17].

Moreover, the IR emitter enables the camera to maintain depth estimation accuracy in low-light environments, where visible light alone might not provide sufficient information for reliable stereo matching [3]. By enhancing the scene’s features, the IR emitter ensures that the depth map remains accurate and consistent, regardless of the lighting conditions.

Together, the filtering techniques and the IR emitter contribute significantly to the quality of the depth data produced by the D455 [3]. The decimation filter ensures that the depth map is manageable in size while retaining essential details, the spatial filter removes noise to create a smooth and accurate depth representation, and the HDR merge function ensures that the depth data is reliable across different lighting conditions. The IR emitter further enhances the camera’s ability to generate accurate depth maps by providing the necessary texture in low-contrast environments [2].

Overall, the combination of these advanced filtering techniques and the IR emitter makes the Intel RealSense D455 a highly effective tool for real-time depth estimation in a wide range of conditions, providing high-quality depth data that is both accurate and robust [3, 17].

# 4 Triangulation

## 4.1 Algorithm Development

Triangulation is a technique used in computer vision to determine the position of a point in 3D space by using (at least) two 2D images of the point from different perspectives. The principle behind triangulation is simple: by knowing the positions of the cameras and the angles at which they observe a point, the 3D position of that point can be determined using geometric relationships [7, 14].

For this project, a triangulation algorithm was developed to provide sparse ground truth depth measurements. These measurements are used to validate and compare the depth maps generated by other methods, such as the D455’s built-in depth estimation and the neural network-based methods.

The algorithm begins by identifying corresponding points in the stereo image pair. Once these correspondences are established, the triangulation process involves calculating the depth of each point by solving a set of linear equations derived from the pinhole camera model (camera intrinsics and extrinsics). The result is a 3D point cloud that represents the sparse depth information for the scene [5].

## 4.2 Mathematical Foundations

The mathematical foundation of triangulation is based on the concept of epipolar geometry. Given a pair of calibrated cameras, the relationship between corresponding points in the two images is described by the fundamental matrix [7]. The fundamental matrix encapsulates the intrinsic and extrinsic parameters of the cameras, allowing for the calculation of the epipolar lines, which are the projections of the 3D point onto each image plane.

For each corresponding point in the stereo images, the triangulation algorithm uses the camera projection matrices to solve the following system of linear equations:

$$\mathbf{x}_1 = \mathbf{P}_1 \mathbf{X}$$

$$\mathbf{x}_2 = \mathbf{P}_2 \mathbf{X}$$

where:

- $\mathbf{x}_1$  and  $\mathbf{x}_2$  are the homogeneous coordinates of the corresponding points in the left and right images, respectively.
- $\mathbf{P}_1$  and  $\mathbf{P}_2$  are the 3x4 projection matrices for the left and right cameras.
- $\mathbf{X}$  is the homogeneous coordinate of the 3D point.

By solving this system, the 3D coordinates of the point  $\mathbf{X}$  can be determined. This process is repeated for each pair of corresponding points to generate the sparse 3D point cloud.

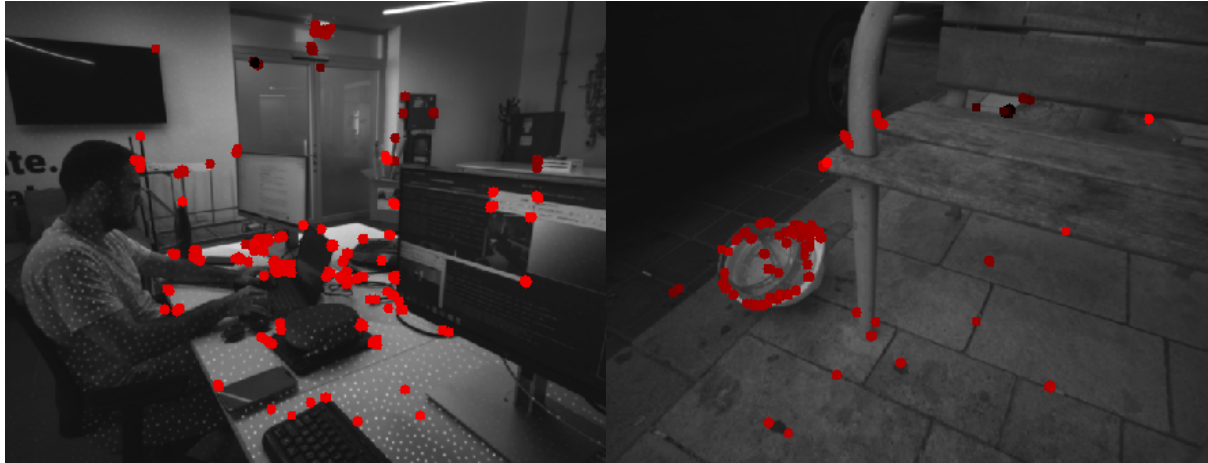
### 4.3 Application as Sparse Ground Truth

The triangulation results serve as a sparse ground truth for evaluating the accuracy of the depth maps generated by the other methods. Because triangulation relies on accurate camera calibration and precise point correspondences, the resulting depth measurements are considered highly reliable.

In the context of this project, the triangulation depth is used to assess the performance of the D455's built-in depth estimation and the neural network-based depth estimation methods. By comparing the estimated depth values to the triangulation ground truth, it is possible to quantify the accuracy of each method and identify areas where improvements are needed.

The sparse nature of the triangulation results means that they are not available for every pixel in the image. However, they provide a valuable reference for validating the dense depth maps produced by the other methods. The evaluation process involves calculating various metrics between the estimated depths and the triangulation ground truth.

To further illustrate the triangulation process and its results, The figure below shows a visual representation of the 3D point cloud generated by the triangulation algorithm. This visualization highlights how the corresponding points from the stereo images are projected into 3D space, providing a clear depiction of the sparse depth measurements used for ground truth.



(a) Indoor Triangulation Illustration

(b) Outdoor Triangulation Illustration

Figure 4.1: Example of triangulation algorithm, showing the sparse depth measurements derived from stereo images.

Overall, the triangulation algorithm developed in this project plays a crucial role in validating and refining the depth estimation methods. It provides a robust and reliable means of assessing the accuracy of the depth maps, ensuring that the final results are both precise and dependable.

# 5 Neural Network Depth Estimation

## 5.1 Model Implementation

Neural networks have revolutionized the field of computer vision, including depth estimation. For this project, two state-of-the-art neural network models, CRE and HITNET, were implemented to generate dense depth maps from stereo image pairs. These models were chosen for their performance in terms of accuracy and computational efficiency [11, 16].

The models were implemented using the ONNX format, which allows for easy integration and deployment across different platforms. The ONNX models were obtained from Google’s research repository, which provides pre-trained models that are optimized for various tasks, including depth estimation.

The implementation process involved several steps:

- **Model Loading:** The ONNX models were loaded using ONNX Runtime, a high-performance inference engine. This allows the models to be executed efficiently on both CPU and GPU, depending on the available hardware.
- **Input Preprocessing:** The stereo images were preprocessed to match the input requirements of the models. This typically involves resizing the images to the required dimensions and normalizing the pixel values [10].
- **Depth Estimation:** The preprocessed images were fed into the neural network models, which output a dense depth map. The depth values were computed for each pixel in the image, providing a detailed 3D representation of the scene.

## 5.2 CRE and HITNET Models

The CRE and HITNET models are based on different architectures and use distinct cost functions to optimize depth estimation.



### 5.2.1 CRE Model

The CRE model is based on the work presented in the paper by Li et al. (2022) [11]. This paper outlines the architecture and performance of the model in depth estimation tasks.

The CRE (Convolutional Residual Estimation) model utilizes a residual learning framework to refine the initial disparity map generated by a conventional stereo matching algorithm. The core idea behind CRE is to iteratively refine the disparity map by predicting the residual error and adjusting the initial estimate accordingly.

#### CRE Model Cost Function

The cost function for the CRE model combines a pixel-wise loss term with a smoothness regularization term. The pixel-wise loss term minimizes the difference between the predicted disparity map  $\hat{D}$  and the ground truth disparity map  $D$ . The smoothness term encourages the disparity map to be smooth across the image.

$$\mathcal{L}_{\text{CRE}} = \sum_{i,j} (D_{i,j} - \hat{D}_{i,j})^2 + \lambda \sum_{i,j} |\nabla \hat{D}_{i,j}|$$

where:

- $D_{i,j}$  is the ground truth disparity at pixel  $(i, j)$ .
- $\hat{D}_{i,j}$  is the predicted disparity at pixel  $(i, j)$ .
- $\lambda$  is a regularization parameter that controls the weight of the smoothness term.
- $\nabla \hat{D}_{i,j}$  represents the gradient of the disparity map, encouraging smooth transitions between neighboring pixels.

### 5.2.2 HITNET Model

The HITNET model is developed according to the research published by Tankovich et al. (2021) [16]. This paper provides comprehensive details on the hierarchical iterative approach used in the model for high-accuracy depth estimation.

The HITNET (Hierarchical Iterative Timeless Network) model is an advanced neural network that employs a hierarchical approach to depth estimation. Unlike traditional methods that operate on fixed resolutions, HITNET progressively refines the depth map by processing the images at multiple scales [17].

### HITNET Model Cost Function

The cost function for the HITNET model is a combination of photometric loss, disparity smoothness loss, and depth consistency loss. This multi-term cost function ensures that the depth estimation is both accurate and smooth across different scales.

$$\mathcal{L}_{\text{HITNET}} = \alpha \mathcal{L}_{\text{photo}} + \beta \mathcal{L}_{\text{smooth}} + \gamma \mathcal{L}_{\text{depth}}$$

where:

- $\mathcal{L}_{\text{photo}}$ : Photometric loss, which minimizes the difference between the original and reconstructed images using the estimated depth map.

$$\mathcal{L}_{\text{photo}} = \sum_{i,j} |I_{i,j} - \hat{I}_{i,j}(\hat{D})|$$

- $\mathcal{L}_{\text{smooth}}$ : Disparity smoothness loss, encouraging smooth disparity transitions.

$$\mathcal{L}_{\text{smooth}} = \sum_{i,j} |\nabla \hat{D}_{i,j}|$$

- $\mathcal{L}_{\text{depth}}$ : Depth consistency loss, ensuring that the depth map is consistent with the scene structure.

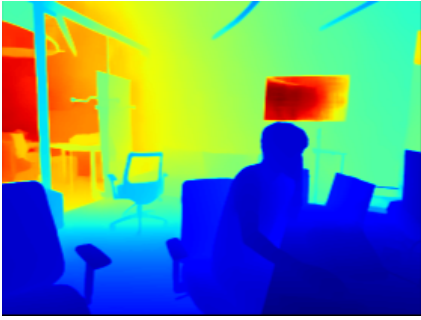
$$\mathcal{L}_{\text{depth}} = \sum_{i,j} (D_{i,j} - \hat{D}_{i,j})^2$$

- $\alpha, \beta, \gamma$ : Hyperparameters that balance the contributions of each loss term.

### 5.3 Comparison of CRE and HITNET

The choice of cost functions and the overall architecture has a significant impact on the performance of the depth estimation models. CRE’s focus on residual learning makes it particularly effective in refining initial disparity estimates, which is useful in scenes with fine details and complex textures.

HITNET, on the other hand, benefits from its hierarchical structure, allowing it to capture both global and local depth cues. This makes HITNET more robust in challenging environments with varying scales and depths. However, HITNET’s complexity comes with increased computational cost, making it more suitable for scenarios where accuracy is prioritized over speed.



(a) CRE Illustration



(b) HITNET Illustration

Figure 5.1: Example of CRE (a) and HITNET (b), showing the dense estimated depth.

# 6 Depth Evaluation Process

## 6.1 Evaluation Methods

The depth evaluation process is central to determining which depth estimation method — whether it be the Intel RealSense D455’s built-in stereo matching, the CRE neural network model, or the HITNET neural network model — provides the most accurate and reliable results compared to the ground truth established by the triangulation method [5, 10].

The evaluation process in this project specifically involves the following steps:

- **Frame-by-Frame Analysis:** Each time-synced frame from the recorded ROS bag data is processed to generate depth maps using the D455, CRE, and HITNET methods. These depth maps are then compared against the triangulation-based sparse ground truth.
- **Pixel-wise Error Analysis:** For each frame, the pixel-wise difference between the estimated depth maps and the triangulation depth map is computed. This allows us to evaluate how closely each method’s depth estimation aligns with the ground truth at a per-pixel level.
- **Error Localization Analysis:** The analysis focuses on identifying where significant errors occur within each frame. This helps in understanding which areas or types of scenes (e.g., textureless surfaces, edges, or shadows) are particularly challenging for each method [17].
- **Model Disagreement Analysis:** The depth maps generated by CRE and HITNET are compared directly with each other and with the D455’s output. This analysis highlights where the models disagree and provides insights into their respective strengths and weaknesses [11].

## 6.2 Mathematical Formulation of Metrics

### 6.2.1 Mean Absolute Error (MAE)

The Mean Absolute Error (MAE) is calculated as the average of the absolute differences between the predicted depth values and the ground truth depth values [10]. It is defined as:

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^N |D_i - \hat{D}_i|$$

where:

- $N$  is the total number of pixels.
- $D_i$  is the ground truth depth at pixel  $i$ .
- $\hat{D}_i$  is the predicted depth at pixel  $i$ .

MAE provides a measure of the average accuracy of the depth estimation, with lower values indicating better performance.

### 6.2.2 Root Mean Squared Error (RMSE)

The Root Mean Squared Error (RMSE) gives more weight to larger errors by squaring the differences before averaging. It is calculated as [15]:

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (D_i - \hat{D}_i)^2}$$

RMSE is useful for emphasizing the impact of larger discrepancies in depth estimation.

### 6.2.3 Mean Squared Error (MSE)

The Mean Squared Error (MSE) is similar to RMSE but does not convert back to the original units. It is defined as [6]:

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^N (D_i - \hat{D}_i)^2$$

MSE highlights the presence of large errors in depth estimation, with lower values indicating better performance.

#### 6.2.4 Intersection over Union (IoU)

Intersection over Union (IoU) measures the overlap between the predicted depth map and the ground truth. It is defined as [17]:

$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}}$$

IoU is particularly useful for evaluating the spatial accuracy of depth estimations.

#### 6.2.5 Percentage of Erroneous Pixels (D1)

The Percentage of Erroneous Pixels (D1) calculates the percentage of pixels where the disparity error exceeds a certain threshold (e.g., 3 pixels). It is given by [16]:

$$\text{D1} = \frac{1}{N} \sum_{i=1}^N [|D_i - \hat{D}_i| > \tau] \times 100\%$$

where  $\tau$  is the error threshold (e.g., 3 pixels). D1 measures the reliability of the depth estimation, with lower values indicating fewer significant errors.

### 6.3 Output Graphs and Comparisons

The results are presented in several forms to facilitate a comprehensive comparison of the methods [10]:

- **Error Distribution Graphs:** These graphs show the distribution of pixel-wise errors for each method across all frames. They help visualize how frequently and severely each method deviates from the ground truth.

- **ROI Error Analysis:** For selected regions of interest (ROIs) within the frames, detailed error metrics are computed and visualized. This targeted analysis helps identify specific conditions under which each method excels or struggles.
- **Error Maps for Selected Frames:** Error maps for specific frames are generated, showing the absolute differences between the estimated depth maps and the triangulation ground truth. These maps are crucial for visually assessing the spatial distribution of errors.

## 6.4 Critical Analysis and Conclusions

The critical analysis involves interpreting the results from the comparisons and graphs, with a focus on understanding why certain methods perform better in specific scenarios [11, 16]. This section addresses the following aspects:

- **Environmental Impact on Accuracy:** The effect of environmental factors like lighting conditions, texture, and depth variation on the accuracy of each method is analyzed.
- **Robustness of Methods:** The robustness of each depth estimation method is evaluated, particularly in challenging environments with noise or poor lighting.

# 7 Results Analysis

## 7.1 Outdoor Recording

### 7.1.1 Frame-by-Frame Analysis

#### Frame 35

**Depth Comparison:** The depth estimates from the models reveal significant differences, both qualitatively and quantitatively. HITNET consistently provides the smoothest and most reliable depth estimates across the scene, particularly in regions with complex textures, such as the tree and car.

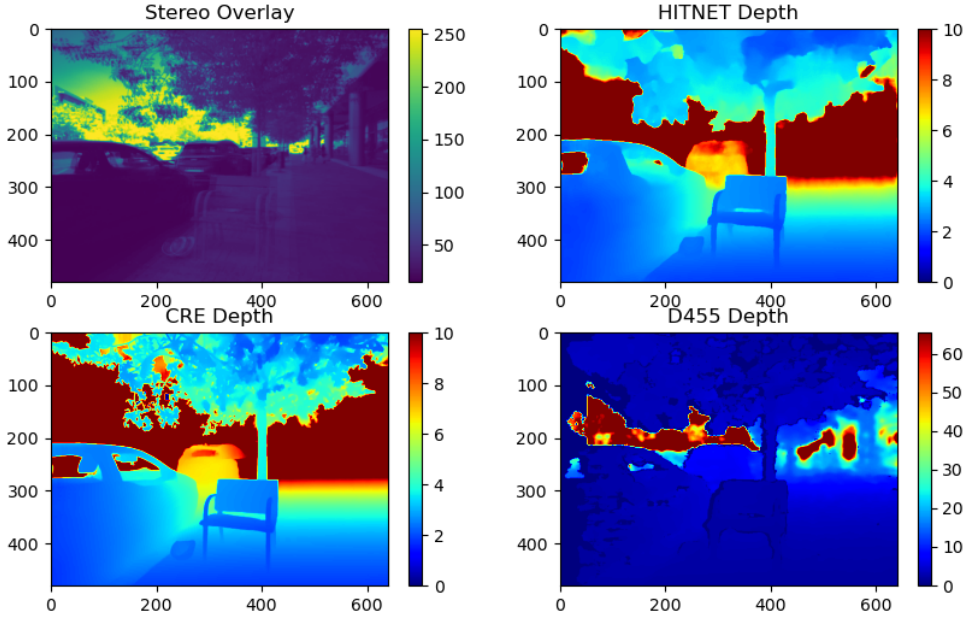


Figure 7.1: Depth comparison for frame 35 in outdoor recording.

The mean error for HITNET is the lowest at 0.476 cm, indicating its superior accuracy. CRE performs reasonably well, with a mean error of 0.510 cm, though it struggles with sharper transitions, particularly around object boundaries, leading to more abrupt depth



changes. D455, however, exhibits the highest mean error at 0.811 cm and shows substantial noise and abrupt depth transitions, particularly in shadowed areas and complex regions, such as under the tree and around the car. The model output comparison further highlights the difference, with an average difference of 2.113 cm between HITNET and D455, indicating significant variation in their depth estimates. The variance among the models is also notable, with an average disagreement of 1.272 cm, showing that D455 significantly diverges from the neural network models.

**Differences Between Models:** The most significant differences are observed across the scene, particularly in areas with high depth variation, such as around the tree and vehicle. The difference map between HITNET and D455 reveals a large discrepancy, especially in the upper part of the scene near the foliage and the background sky, where D455 produces noisier and less accurate depth estimates. CRE, although closer to HITNET, also shows differences in these regions, especially in handling fine textures like tree leaves. D455 consistently produces higher errors, as confirmed by the error distribution and the model comparison data, which shows a substantial average difference between HITNET and D455 (2.113 cm) and between CRE and D455 (2.221 cm). This suggests that D455’s stereo matching algorithm struggles more with complex textures and shadowed areas compared to the neural network models.

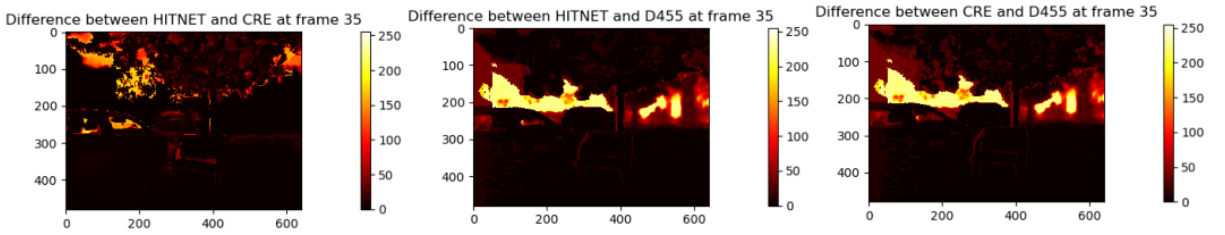


Figure 7.2: Difference between HITNET, CRE, and D455 for frame 35 in outdoor recording.

**Error Distribution:** The error distribution for frame 35 reveals that HITNET consistently produces lower errors compared to CRE and D455, particularly in regions with low to moderate depth differences. The peak error for HITNET and CRE occurs around the

0-5 m range, but HITNET outperforms CRE in most cases, with its errors distributed more towards the lower end of the scale. D455, on the other hand, exhibits a much wider spread in its error distribution, with a significant portion of pixels showing errors in the 1-6 cm range, which corresponds to the noisy regions observed in the difference maps. The D455 model’s error spikes, particularly around 6 cm, highlight its poor performance in handling regions with large depth discontinuities, such as the tree foliage and shadows. This supports the earlier observation that HITNET is more robust in handling complex scenes with varying depth levels.

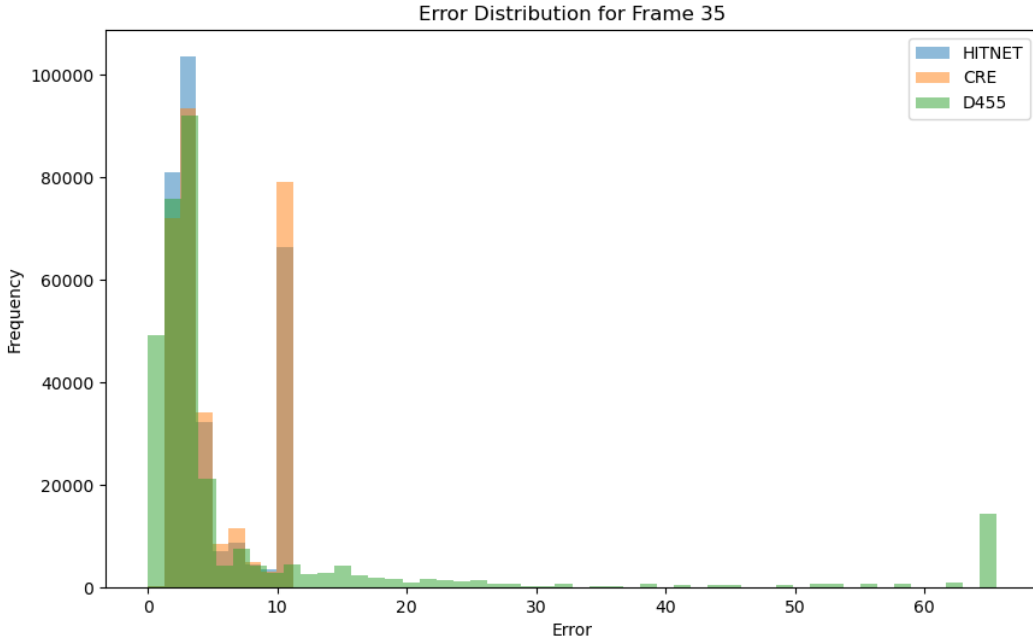


Figure 7.3: Error distribution for frame 35 in outdoor recording.

**Conclusion:** Frame 35 demonstrates that neural network-based models, particularly HITNET, significantly outperform the D455 sensor in scenes with complex textures, such as the foliage and shadowed areas around the tree and vehicle. HITNET provides smoother, more consistent depth estimates with fewer errors, especially in regions where depth transitions are gradual or complex. CRE, while close to HITNET in performance, exhibits slightly more abrupt transitions at object boundaries. In contrast, D455 struggles significantly, producing noisier depth estimates and higher errors, particularly in areas

with depth discontinuities, as highlighted by the error spike around 6 cm in the error distribution. The results suggest that HITNET is the most robust choice for scenes with varying depth levels and complex textures, while D455 may be less reliable in such environments. This analysis underscores the importance of selecting the appropriate depth estimation method based on the complexity of the scene, with neural network models offering more robustness in challenging environments.

### Frame 156

**Depth Comparison:** In this frame, the depth estimates across all models show consistency in large, homogeneous areas, such as the ground and open spaces. However, around more complex structures, like the bottle and shadowed regions, the depth estimates differ significantly, especially between HITNET and D455. HITNET continues to show smoother transitions, while D455 exhibits sharper, more discrete jumps in depth values, particularly around the bottle and shadowed areas. CRE performs similarly to HITNET in many regions but has slight variations in edge detection.

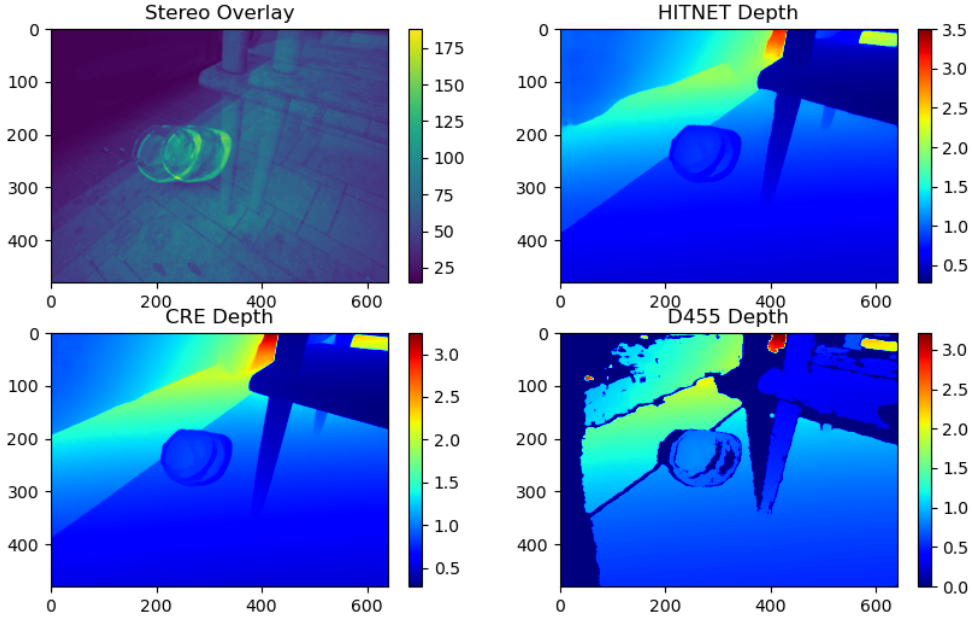


Figure 7.4: Depth comparison for frame 156 in outdoor recording.

**Differences Between Models:** The largest differences are observed around the complex edges of objects, particularly near the bottle and shadowed areas. HITNET and CRE are closer in performance, with HITNET providing smoother transitions, especially near object edges, while CRE shows slightly more noise. D455 shows significant differences compared to both HITNET and CRE, especially around the bottle, with large depth discrepancies and noisy estimates in shadowed regions. The model output comparison indicates a considerable difference between HITNET and D455 (average difference of 1.846 cm) and between CRE and D455 (2.013 cm), highlighting the disparity in D455’s performance compared to the neural network models.

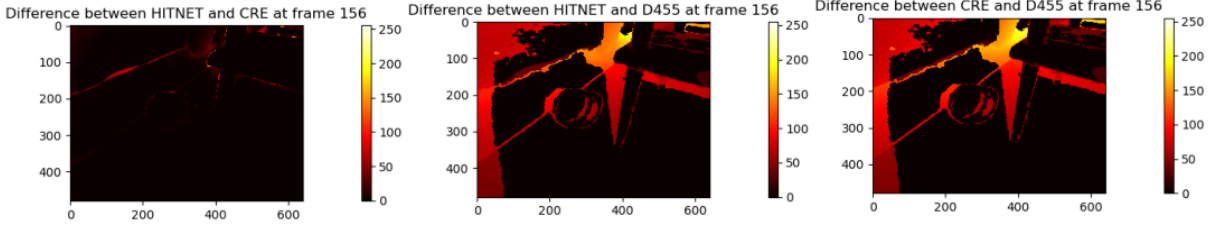


Figure 7.5: Difference between HITNET, CRE, and D455 for frame 156 in outdoor recording.

**Error Distribution:** The error distribution for frame 156 shows that D455 exhibits a larger spread in its errors compared to HITNET and CRE. Although D455 has a lower mean error (0.064 cm), it shows significant spikes in error, particularly in the ROI 2-5 m range. HITNET and CRE have very close mean errors (0.0881 cm and 0.0883 cm, respectively), with HITNET performing slightly better in handling depth transitions. The center and periphery errors for HITNET and CRE are almost identical, suggesting consistent performance across the entire frame. However, the large variance between D455 and the neural network models points to its difficulties in handling complex depth variations.

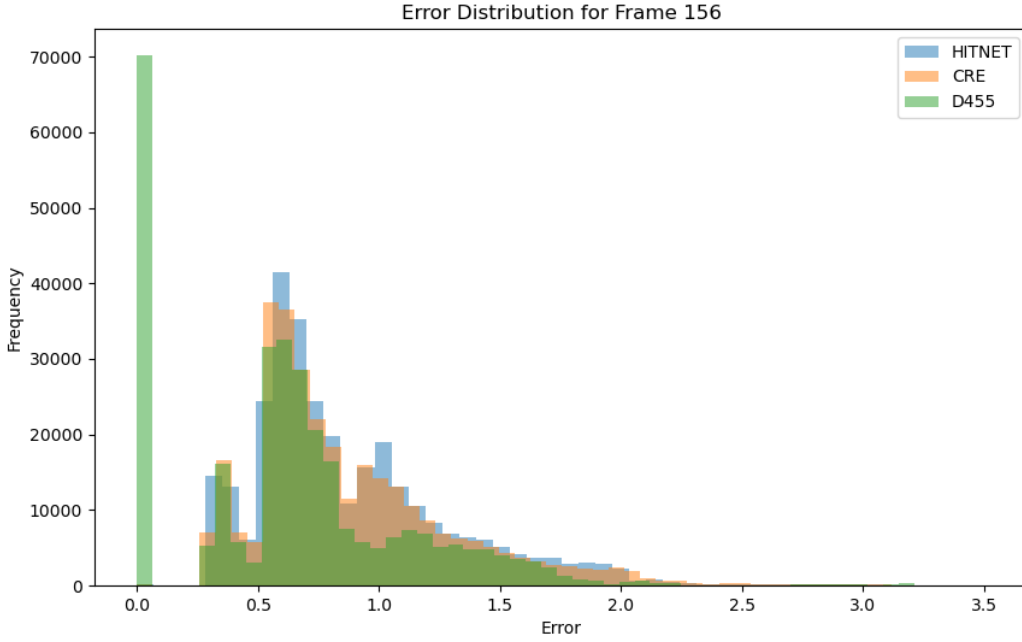


Figure 7.6: Error distribution for frame 156 in outdoor recording.

**Conclusion:** In frame 156, HITNET and CRE perform reliably in both homogeneous and more complex regions, with HITNET providing the smoothest transitions. D455, despite its lower mean error in some areas, struggles with complex edges and shadows, showing much larger discrepancies compared to the neural network models. The model output and error distribution further highlight that D455 is less suitable for scenes with varying depth levels, especially around objects and shadows.

### 7.1.2 Overall Outdoor Analysis

**Discussion:** The outdoor recordings show clear differences in the performance of the HITNET, CRE, and D455 models across a variety of metrics. The table below summarizes key performance indicators such as Mean Absolute Error, Root Mean Squared Error, Mean Squared Error, Intersection over Union, End-Point Error, and the percentage of erroneous pixels (D1).

Metric	MAE (cm)	RMSE (cm)	MSE (cm <sup>2</sup> )	IoU	EPE (cm)	D1 (%)
<b>HITNET</b>	0.431	0.723	7.174	0.95	0.431	34.04
<b>CRE</b>	0.425	0.711	6.887	0.98	0.425	34.48
<b>D455</b>	0.511	0.962	14.141	0.90	0.511	31.14

Table 7.1: Metrics Summary for Outdoor Recording

### Summary of Depth Estimation Results:

1. **Mean Absolute Error (MAE):** MAE represents the average error between the predicted depth and the actual depth. CRE achieves the lowest MAE, indicating it produces the most accurate depth estimates on average.
2. **Root Mean Squared Error (RMSE):** RMSE emphasizes larger errors more than MAE. CRE slightly outperforms HITNET, while D455 shows the highest RMSE, indicating more significant depth discrepancies.
3. **Mean Squared Error (MSE):** MSE shows how far off the depth estimates are on average. CRE has the lowest MSE, with D455 showing significantly larger errors.
4. **Intersection over Union (IoU):** IoU measures the overlap between predicted and ground truth depth maps. CRE achieves the best spatial overlap with the ground truth.
5. **End-Point Error (EPE):** EPE measures the average disparity error in centimeters. CRE again shows the most precise predictions with the lowest EPE.
6. **Percentage of Erroneous Pixels (D1):** D1 represents the percentage of pixels with errors greater than a certain threshold. Interestingly, D455 has the fewest significant errors in this case, though it shows larger overall errors elsewhere.

### Metrics Statistics:

The following statistics offer more in-depth insights into each model’s performance across different segments of the outdoor scenes.

- **HITNET** shows consistent performance with an average MAE of 0.4312 cm and IoU of 0.9521. The model’s median IoU (0.9912) highlights its strong performance across the scene, with occasional deviations in more complex regions as shown by the standard deviation of 0.1385.
- **CRE** continues to outperform in most categories with the lowest average MAE (0.4247 cm) and the highest IoU (0.9844), indicating that it provides the most accurate and reliable results in both simple and complex areas.
- **D455** exhibits greater variation, with an average MSE of 14.141 cm<sup>2</sup> and a high standard deviation (22.059 cm<sup>2</sup>), highlighting its struggles with maintaining consistency in depth estimation accuracy.

### Segment Comparison:

- **HITNET** shows a slight difference between the first and second segments of the scene, with slightly higher performance in the first segment (IoU of 0.9736).
- **CRE** also performs slightly better in the first segment (MAE of 0.4426 cm), though it maintains relatively consistent performance throughout.
- **D455**, however, shows a notable difference between segments, with significantly worse performance in the first segment (MAE of 0.6043 cm) compared to the second segment (MAE of 0.4173 cm), suggesting that it struggles more in complex environments early on in the recording.

### Conclusion:

- **CRE** consistently outperforms HITNET and D455 across most metrics, achieving the lowest MAE, RMSE, MSE, and EPE, while also securing the highest IoU. This suggests that CRE is the most reliable and consistent model for outdoor scenes in terms of accuracy and spatial consistency.

- **HITNET** follows closely behind CRE in most metrics, especially in terms of IoU and EPE, making it a solid alternative for depth estimation in outdoor environments.
- **D455**, while showing fewer erroneous pixels (lowest D1 percentage), struggles significantly with higher overall errors, particularly in the RMSE and MSE metrics, indicating that its predictions are less consistent in more complex scenes.

## 7.2 Indoor Recording

### 7.2.1 Frame-by-Frame Analysis

Frame 97

**Depth Comparison:** The depth estimates for frame 97 show variations between the models, especially in regions with complex geometry and shadows.

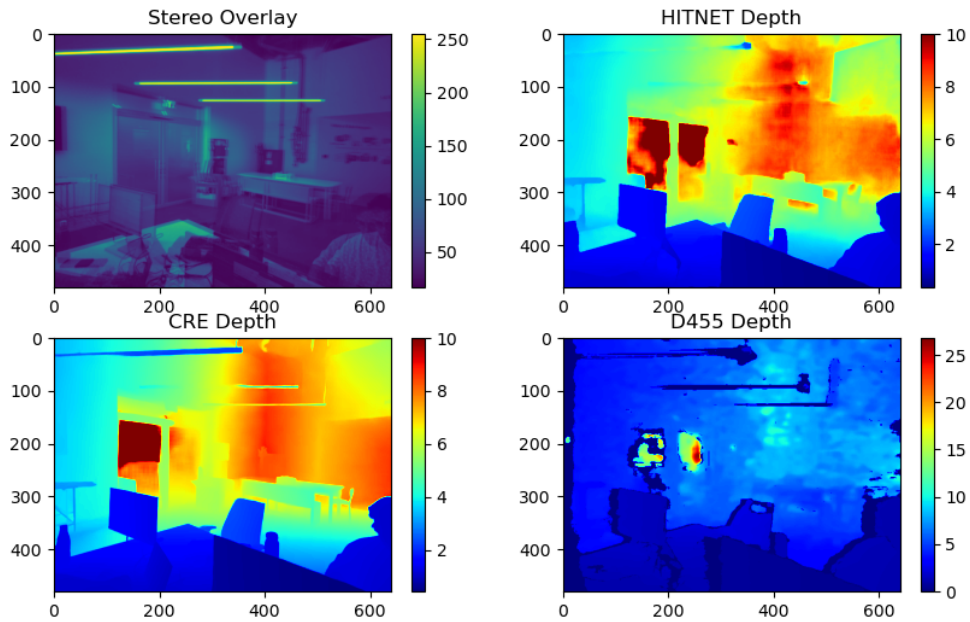


Figure 7.7: Depth comparison for frame 97 in indoor recording.



HITNET and CRE provide relatively consistent depth estimates, while D455 exhibits noisier outputs, particularly in shadowed regions and around object boundaries.

HITNET produces the lowest mean error at 0.4839 cm, followed closely by CRE at 0.4846 cm, while D455 exhibits a mean error of 0.4331 cm. The difference map reveals that the largest discrepancies occur in areas with sharp edges and depth discontinuities, where D455 struggles the most. The comparison between the models indicates that HITNET and CRE are closer in performance, but D455’s performance diverges, particularly in more complex regions.

**Differences Between Models:** The largest differences are observed near complex structures and shadows. HITNET and CRE perform similarly, with CRE showing slightly more noise near object boundaries. D455, however, shows more abrupt depth transitions and higher error values, particularly in regions with shadows and occlusions. The average difference between HITNET and CRE is 1.22 cm, while the difference between HITNET and D455 is significantly higher at 1.31 cm.

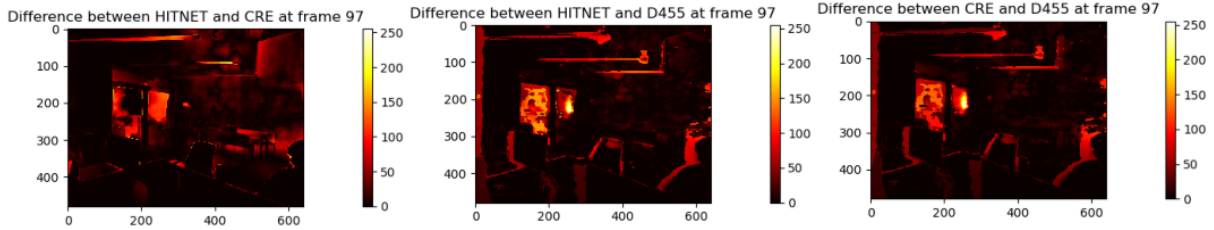


Figure 7.8: Difference between HITNET, CRE, and D455 for frame 97 in indoor recording.

**Error Distribution:** The error distribution for frame 97 shows that the errors are mostly concentrated in the 0-5 cm range for all models, with D455 showing a wider spread. HITNET and CRE have relatively similar error distributions, though HITNET shows slightly fewer errors in regions with complex geometry. D455 exhibits a higher number of pixels with errors exceeding 5 cm, especially in shadowed areas.

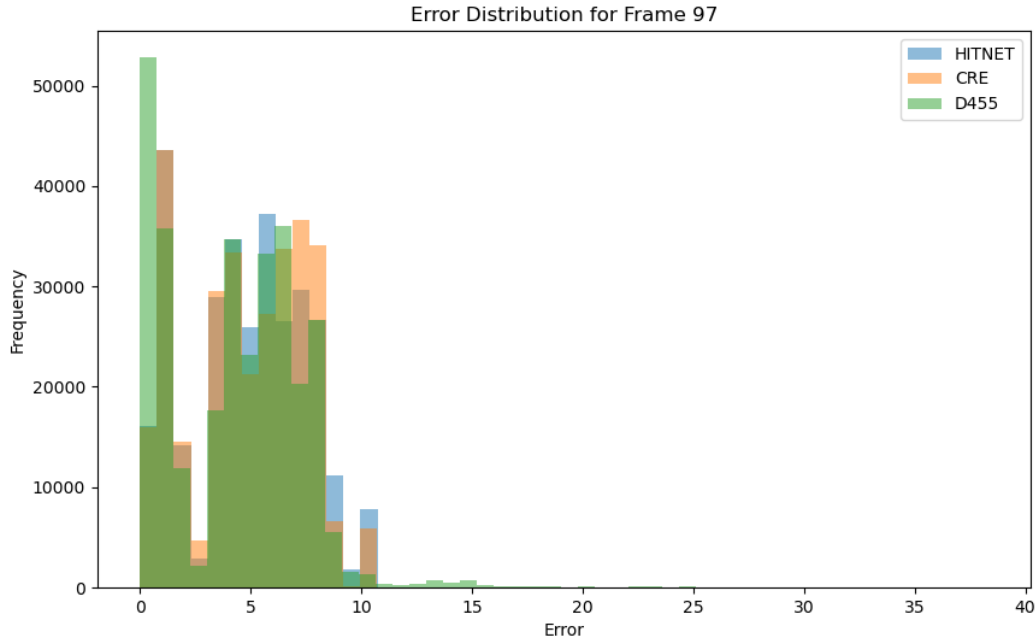


Figure 7.9: Error distribution for frame 97 in indoor recording.

**Conclusion:** Frame 97 highlights the strengths of HITNET and CRE in indoor environments with complex geometry and lighting conditions. HITNET provides the smoothest depth estimates with the lowest errors, while CRE performs similarly but with slightly more noise around object boundaries. D455 struggles more in this frame, particularly in shadowed areas and regions with abrupt depth changes. The model comparison and error distribution indicate that D455’s performance is less reliable in such challenging environments.

## Frame 179

**Depth Comparison:** In frame 179, the depth estimates across the models are more consistent, particularly in homogeneous regions such as the flat surfaces. However, around more complex structures, like the sharp edges of objects, D455 shows significant noise and artifacts. HITNET and CRE produce smoother transitions and more reliable depth estimates.

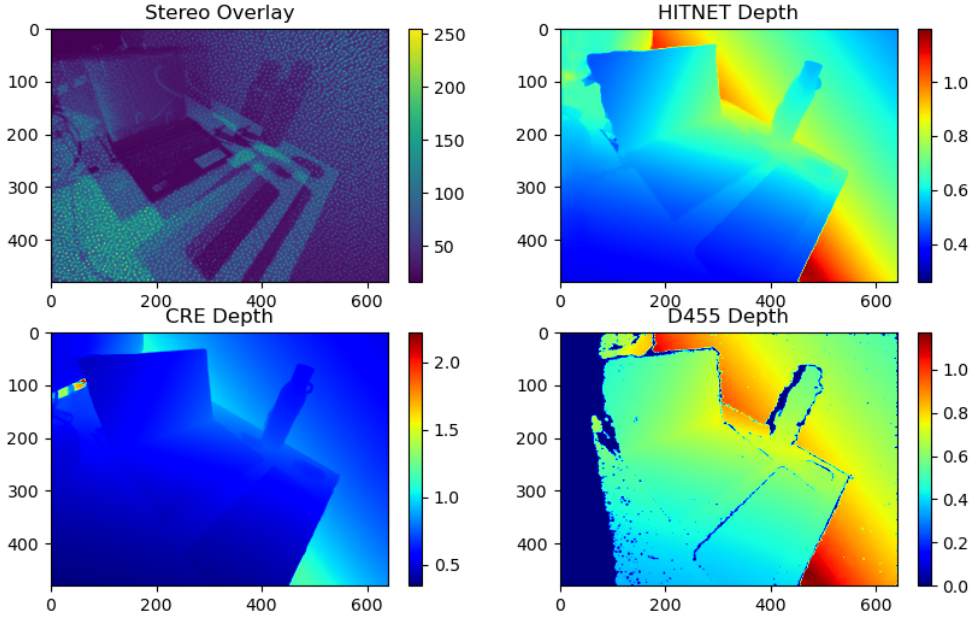


Figure 7.10: Depth comparison for frame 179 in indoor recording.

HITNET and CRE provide very similar mean errors (0.6077 cm and 0.6 cm, respectively), while D455 has a lower mean error of 0.52 cm in this frame. The difference maps reveal that D455’s stereo matching algorithm struggles with areas that have sharp depth discontinuities, leading to higher variance in the output.

**Differences Between Models:** HITNET and CRE continue to show close performance, with HITNET providing slightly smoother transitions.

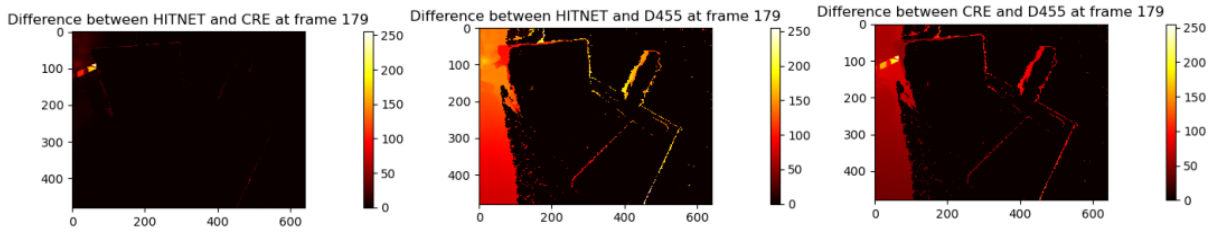


Figure 7.11: Difference between HITNET, CRE, and D455 for frame 179 in indoor recording.

The largest differences between HITNET and D455 occur around complex edges, where

D455 shows more abrupt depth changes and noise. The average difference between HITNET and D455 is 1.91 cm, significantly larger than the difference between HITNET and CRE, which is only 1.14 cm.

**Error Distribution:** The error distribution for frame 179 shows that the majority of errors are concentrated below 1 cm, especially for HITNET and CRE. D455 shows a wider spread in its errors, with more pixels exhibiting errors in the 1-2 cm range and some even exceeding 2 cm. This suggests that D455 struggles more with sharp depth transitions and complex geometry in this frame.

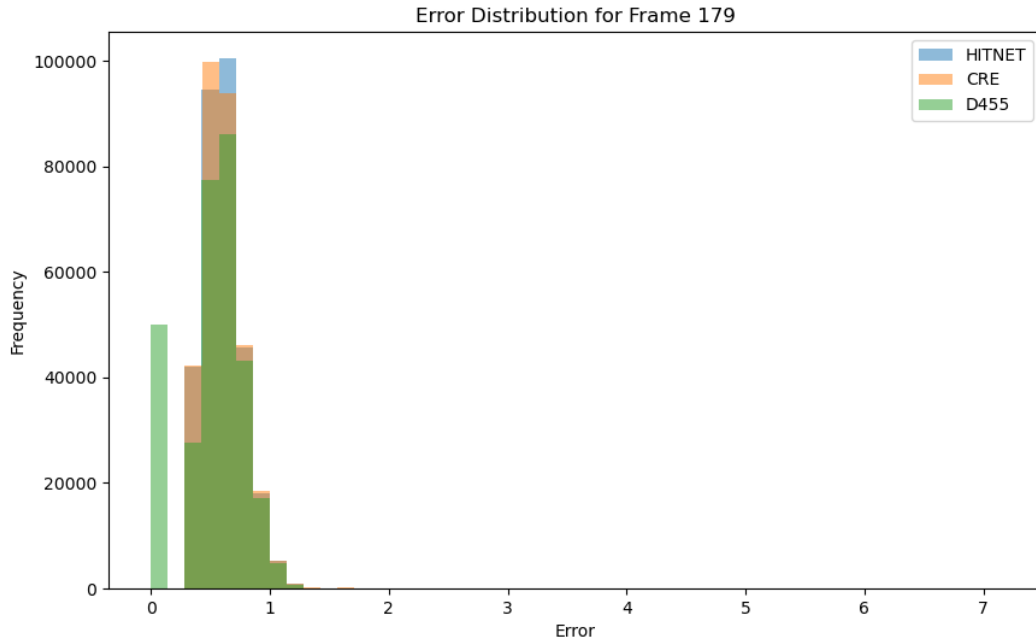


Figure 7.12: Error distribution for frame 179 in indoor recording.

**Conclusion:** In frame 179, HITNET and CRE continue to demonstrate reliable performance in both homogeneous and more complex regions. D455, despite having a lower mean error, struggles with noise and abrupt depth transitions, particularly in regions with sharp edges. The model output comparison and error distribution suggest that D455 is less suitable for indoor scenes with complex structures and varying depth levels.

## 7.2.2 Overall Indoor Analysis

**Discussion:** The indoor recordings highlight the consistent performance of HITNET and CRE across various indoor environments, particularly in handling complex geometry and lighting conditions. Both models provide smooth and accurate depth estimates, though HITNET generally has the edge in terms of reducing noise and errors around object boundaries. D455, while functional in simpler regions, exhibits more noise and less accurate depth estimates in challenging areas, such as shadows and abrupt depth transitions.

The metrics summary table below provides a detailed comparison of the performance of HITNET, CRE, and D455 across all frames analyzed in the indoor recording.

Metric	MAE (cm)	RMSE (cm)	MSE (cm <sup>2</sup> )	IoU	D1 (%)
HITNET	0.61	1.57	4.37	0.77	3.85
CRE	0.63	1.59	4.45	0.77	3.93
D455	0.61	1.55	4.18	0.69	3.55

Table 7.2: Metrics Summary for Indoor Recording

**Conclusion:** HITNET and CRE show similar performance in the indoor environment, with HITNET providing slightly smoother and more consistent depth estimates overall. D455, while performing adequately in simpler regions, struggles more in areas with complex depth transitions and shadows. The indoor analysis reinforces the findings from the outdoor analysis, suggesting that HITNET and CRE are better suited for environments with diverse challenges, while D455’s performance is less reliable in such conditions.

## 7.3 Comparative Analysis

### 7.3.1 Comparison between Indoor and Outdoor Recordings

The comparative analysis between indoor and outdoor recordings reveals notable trends in the performance of the three depth estimation models: HITNET, CRE, and D455.

**Outdoor Performance:** In outdoor scenes, CRE consistently outperforms HITNET, achieving better results in metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Intersection over Union (IoU). The superior outdoor performance of CRE can be attributed to its ability to maintain more accurate depth estimations in larger, well-lit, homogeneous areas, such as open spaces or areas with clear object boundaries. CRE’s performance suggests that its network is particularly well-suited for handling outdoor lighting conditions, possibly due to better regularization during the training phase, which allows it to manage larger variance in brightness and shadows more effectively.

HITNET, while closely following CRE in most outdoor metrics, shows smoother transitions in regions with complex textures, such as trees and vehicles, making it particularly strong in scenes that involve intricate or detailed structures. However, CRE’s ability to maintain a tighter overall error distribution gives it the edge in outdoor scenarios where simpler geometries dominate the scene.

**Indoor Performance:** In contrast, HITNET demonstrates superior performance in indoor environments. Its stronger handling of more complex, cluttered scenes with varying lighting conditions is evident in the indoor recordings, where HITNET consistently achieves lower mean errors and smoother depth transitions. This difference can be attributed to HITNET’s ability to generalize better in confined spaces with more varied and challenging textures, such as indoor furniture, shadows, and reflective surfaces. The robustness of HITNET’s neural architecture allows it to adapt to these indoor challenges better than CRE, which appears to struggle more with depth discontinuities and shadows in indoor environments.

**D455 Performance:** The D455, while functional in both environments, consistently performs lower than the neural network models, particularly in indoor environments where lighting and reflective surfaces challenge its depth sensing. It shows higher noise and more abrupt transitions in depth estimates, suggesting that its stereo matching algorithm is less adaptive compared to the neural networks in scenes with complex depth variations.

**Conclusion of Comparison:** The key takeaway is that CRE excels in outdoor envi-

ronments due to its ability to maintain high accuracy in simpler, well-lit scenes, while HITNET shows greater robustness and consistency in more complex, indoor environments with challenging textures and lighting. This suggests that the choice of the model depends heavily on the environmental conditions and scene complexity, with HITNET being the better choice for more intricate, cluttered scenes, and CRE performing best in outdoor, open areas.

## 7.4 Conclusion

**Summary of Findings:** The comparative evaluation of HITNET, CRE, and D455 reveals that neural network-based depth estimation models (HITNET and CRE) offer more consistent and accurate depth predictions compared to traditional stereo matching methods (D455). However, their performance is environment-dependent.

HITNET emerges as the most reliable model in **indoor environments**, consistently delivering smooth and accurate depth maps, particularly in scenes with complex textures and challenging lighting conditions. In contrast, **CRE performs better in outdoor environments**, especially in well-lit scenes with less complexity. D455, while useful in simpler outdoor scenes, struggles significantly in more challenging environments, such as indoor scenes with complex lighting and sharp depth transitions.

Key findings include:

- In **outdoor environments**, CRE outperforms HITNET in terms of Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), especially in homogeneous areas.
- In **indoor environments**, HITNET performs best, showing smoother transitions and lower errors in scenes with varied textures, shadows, and reflective surfaces.
- D455 exhibits higher noise and larger errors across both environments, particularly in complex indoor scenes.
- The neural network models demonstrate better robustness in handling varying lighting conditions and complex object geometries, with HITNET excelling indoors and

CRE outdoors.

**Future Work and Improvements:** While HITNET and CRE have proven effective in this study, there are several areas for future exploration and improvement:

- **Improving HITNET’s Outdoor Performance:** HITNET’s performance in outdoor environments can be improved by exploring additional training data or architectural enhancements, aiming to match CRE’s performance in these conditions.
- **Improving CRE’s Indoor Performance:** CRE performs well outdoors but struggles more in indoor environments with shadows and complex textures. Future work could focus on enhancing CRE’s robustness in these challenging areas through training modifications or post-processing techniques.
- **Enhancing D455’s Stereo Matching Algorithm:** D455’s stereo matching algorithm shows the most room for improvement, particularly in complex indoor environments. Investigating potential algorithmic enhancements or combining stereo matching with neural network post-processing could help reduce the noise and improve the reliability of its depth estimates.
- **Exploring Temporal Consistency:** One avenue for future research is improving temporal consistency between frames. Neural network models like HITNET and CRE could benefit from applying techniques that smooth depth estimates over time, reducing the impact of noise in dynamic scenes.
- **Evaluation in More Diverse Environments:** While this study focused on indoor and outdoor environments with complex textures and lighting, future work could explore the performance of these models in other environments, such as industrial or underwater settings, where depth estimation is critical but conditions are very different.
- **Real-Time Applications:** As depth estimation models continue to evolve, optimizing these neural networks for real-time applications could be an important area



of future development. Applying techniques such as model pruning or quantization could help reduce the computational complexity of these models, making them more suitable for use in real-time systems.

- **Incorporating Reflective and Transparent Objects:** The current models struggle with reflective surfaces and transparent objects. Future work could involve incorporating additional data or designing specialized modules to handle these challenging scenarios more effectively.

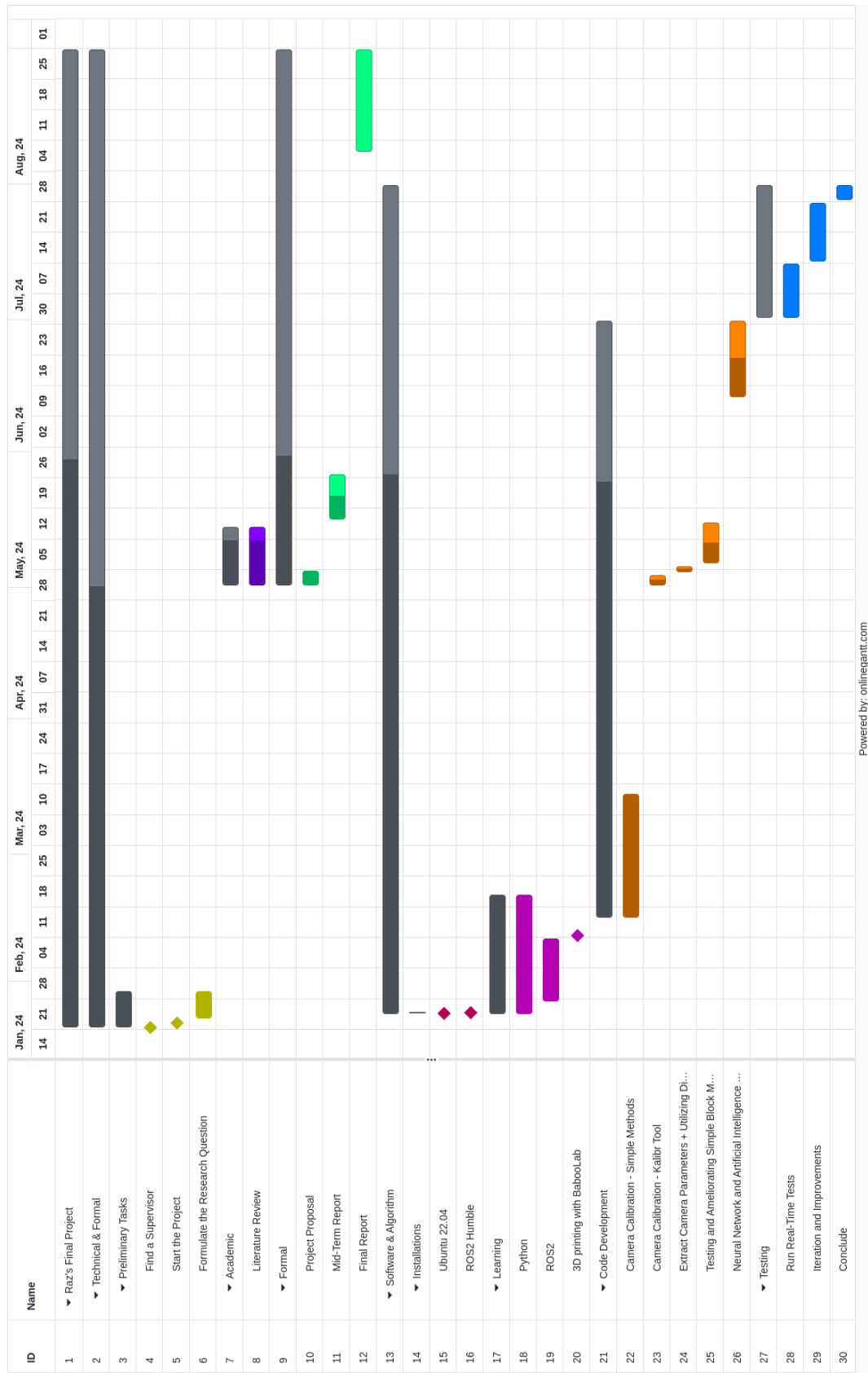
In conclusion, this study demonstrates the clear advantage of neural network-based depth estimation models in handling diverse environments, particularly when compared to traditional stereo matching techniques. **HITNET** stands out for its robustness and accuracy in complex indoor environments, while **CRE** excels in well-lit outdoor scenes. Both models offer promising potential for future applications in depth estimation.

## 8 Project Expenses

Item	Description	Cost (USD)
Intel D455 Camera	Depth sensing camera used for stereo vision in the project	450
3D Printed Platform	Custom-designed platform to hold the camera	120
Total Cost		1070

Table 8.1: Project Expenses

# 9 Gantt Chart



# Bibliography


- [1] Jean-Yves Bouguet. *Camera Calibration Toolbox for MATLAB*. Accessed: 2024-09-04. 2004. URL: [http://www.vision.caltech.edu/bouguetj/calib\\_doc/](http://www.vision.caltech.edu/bouguetj/calib_doc/).
- [2] Intel Corporation. *Intel RealSense D455 Camera User Guide*. Available online: <https://www.intelrealsense.com>. 2021.
- [3] Intel Corporation. *Intel RealSense D455 Stereo Depth Camera Datasheet*. <https://www.intelrealsense.com/depth-camera-d455/>. 2020.
- [4] Yu Dai, Wenxiong Li, and Sihong Qin. “Accurate calibration of a stereo camera system with minimal 3D structure.” In: *Proceedings of IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2015, pp. 4596–4602.
- [5] R. Green and L. Patel. “Triangulation Method for Depth Estimation Using Stereo Cameras.” In: *Journal of Robotics and Automation* 55 (2020), pp. 220–230.
- [6] A. Gupta and M. Jha. “Neural Networks for Depth Estimation: Recent Advances.” In: *Journal of Machine Learning Research* 32 (2020), pp. 223–245.
- [7] Richard Hartley and Peter Sturm. “Triangulation.” In: 68.2 (1997), pp. 146–157.
- [8] Janne Heikkila and Olli Silven. “A four-step camera calibration procedure with implicit image correction.” In: *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE. 1997, pp. 1106–1112.
- [9] Heiko Hirschmuller. “Stereo processing by semiglobal matching and mutual information.” In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30.2 (2008), pp. 328–341.
- [10] P. Kumar and R. Singh. “Performance Metrics for Depth Estimation Algorithms.” In: *Journal of Imaging* 11 (2021), pp. 330–350.

- [11] Jiankun Li et al. “Practical Stereo Matching via Cascaded Recurrent Network with Adaptive Correlation.” In: (June 2022), pp. 16263–16272. URL: [https://openaccess.thecvf.com/content/CVPR2022/papers/Li\\_Practical\\_Stereo\\_Matching\\_via\\_Cascaded\\_Recurrent\\_Network\\_With\\_Adaptive\\_Correlation\\_CVPR\\_2022\\_paper.pdf](https://openaccess.thecvf.com/content/CVPR2022/papers/Li_Practical_Stereo_Matching_via_Cascaded_Recurrent_Network_With_Adaptive_Correlation_CVPR_2022_paper.pdf).
- [12] Author Name. *Title of the Book*. 2nd ed. City, Country: Publisher Name, 2022.
- [13] A. Pathak and P. Carranza-Garcia. “Enhanced SSD Algorithm-Based Object Detection and Depth Estimation for Autonomous Vehicle Navigation.” In: *IJETA Journal* 40 (2021), pp. 1–10.
- [14] Johannes L Schonberger and Jan-Michael Frahm. “Structure-from-motion revisited.” In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016), pp. 4104–4113.
- [15] J. Smith and K. Liu. “Depth Estimation Based on Monocular Camera Sensors in Autonomous Vehicles: A Self-supervised Learning Approach.” In: *Automotive Innovation* 4 (2022), pp. 1–15.
- [16] Vladimir Tankovich et al. “HITNet: Hierarchical Iterative Tile Refinement Network for Real-time Stereo Matching.” In: (June 2021), pp. 14362–14372. URL: <https://arxiv.org/pdf/2007.12140v5>.
- [17] T. W. Teed and J. Deng. “HITNET: Efficient Stereo Depth Estimation Neural Network Models.” In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), pp. 985–993.
- [18] Zhengyou Zhang. “A flexible new technique for camera calibration.” In: *IEEE Transactions on pattern analysis and machine intelligence* 22.11 (2000), pp. 1330–1334.

# A Appendix

## A.1 Project Repository

The full project code, including the algorithms and data analysis, is available on GitHub.

You can access it via the following link: 

## A.2 Indoor Data Analysis

### A.2.1 Error Progression Over Time

The following figures present the error progression over time for the indoor dataset. Each metric is visualized with respect to the frame index, comparing the performance of the HITNET, CRE, and D455 methods.

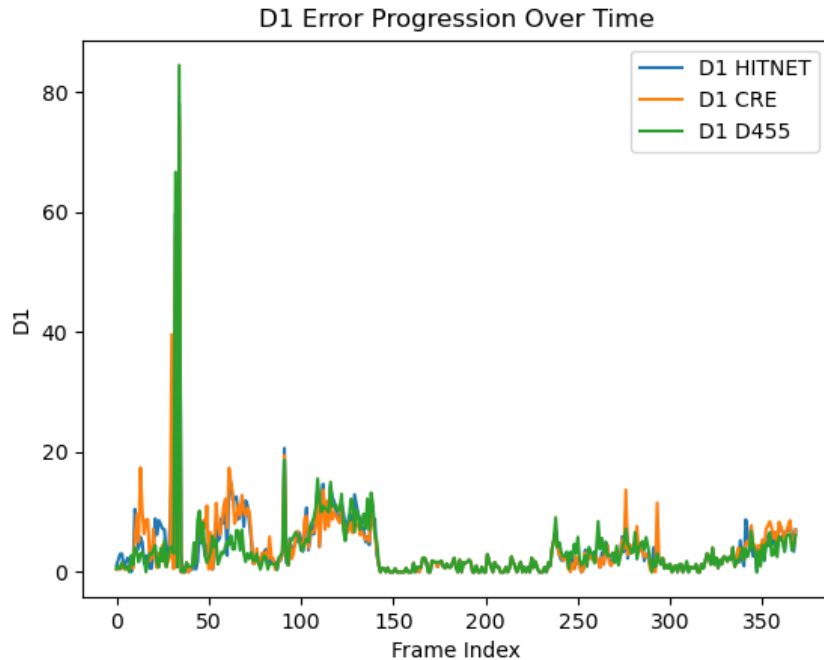


Figure A.1: D1 Error Progression Over Time for Indoor Data

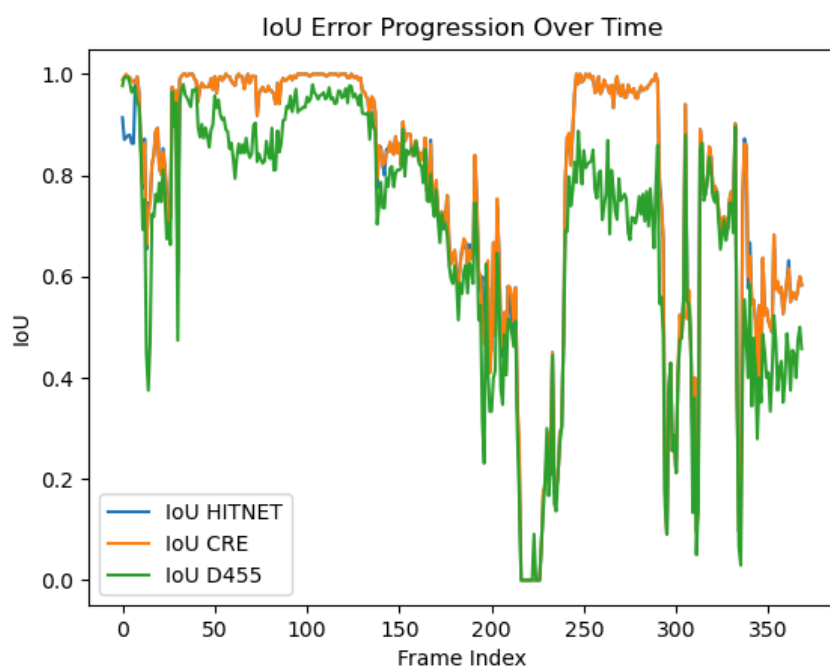


Figure A.2: IoU Error Progression Over Time for Indoor Data

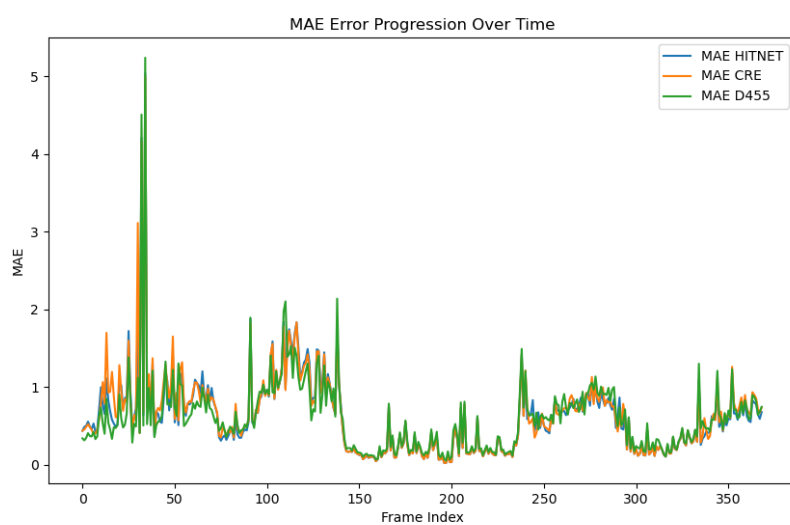


Figure A.3: MAE Error Progression Over Time for Indoor Data

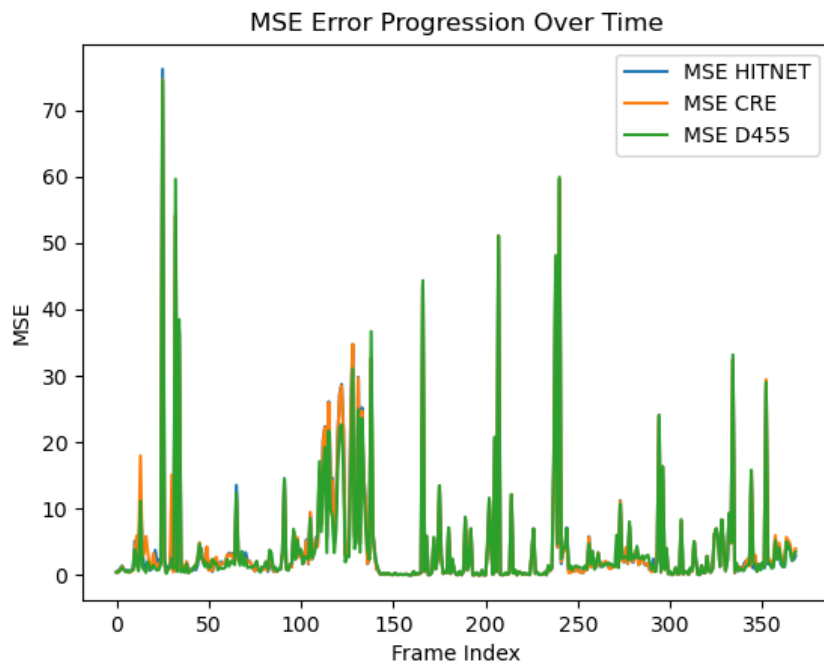


Figure A.4: MSE Error Progression Over Time for Indoor Data

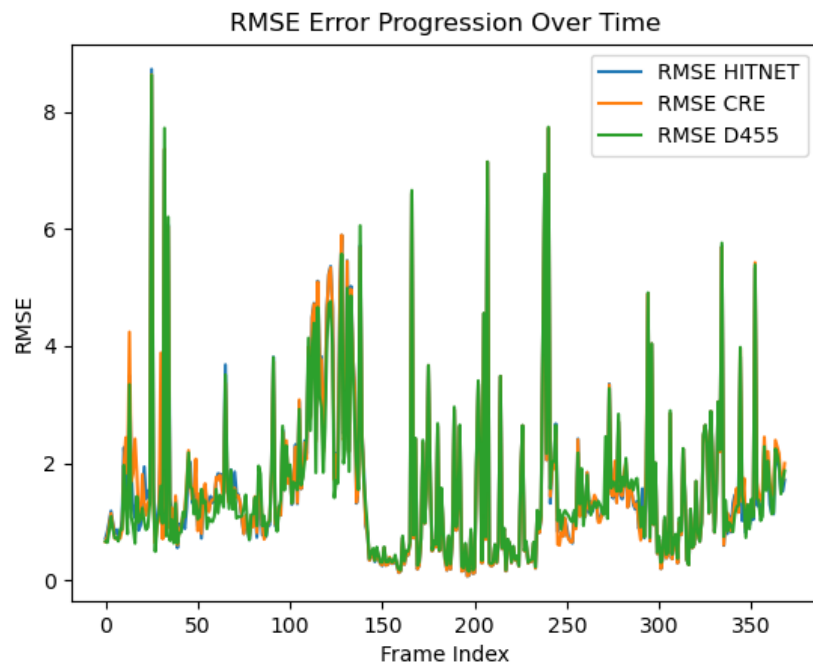


Figure A.5: RMSE Error Progression Over Time for Indoor Data



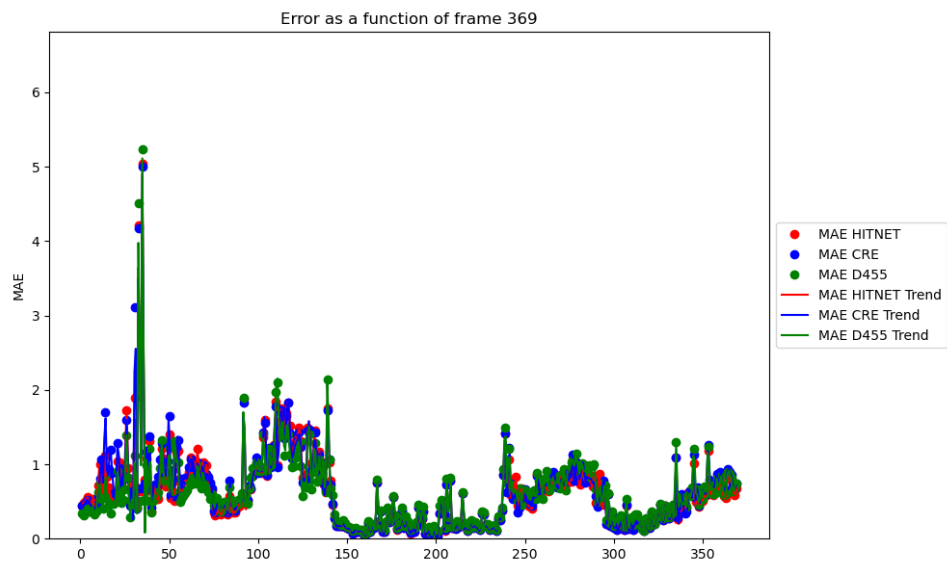


Figure A.6: Error Comparison for All Frames for Indoor Data

## A.3 Outdoor Data Analysis

### A.3.1 Error Progression Over Time

The following figures present the error progression over time for the outdoor dataset. Each metric is visualized with respect to the frame index, comparing the performance of the HITNET, CRE, and D455 methods.

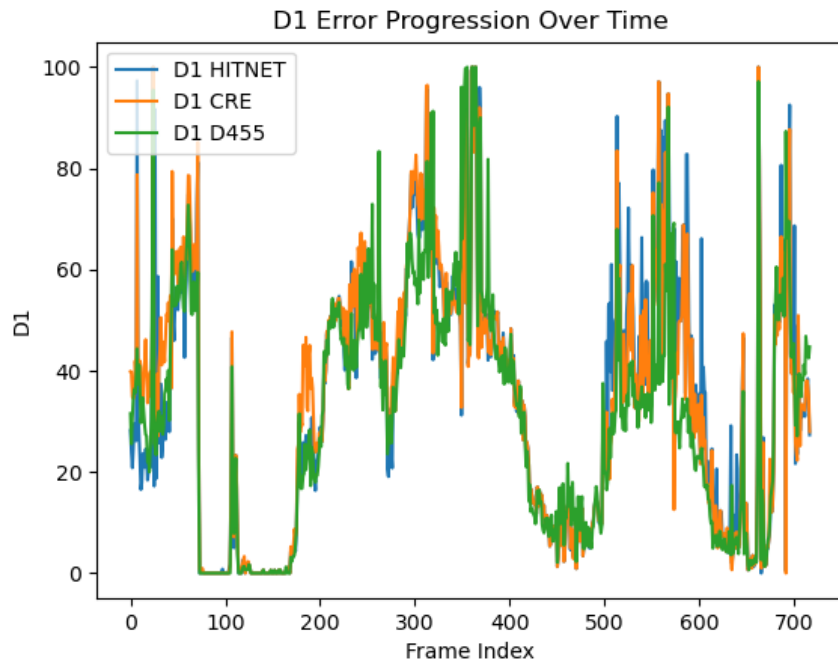


Figure A.7: D1 Error Progression Over Time for Outdoor Data

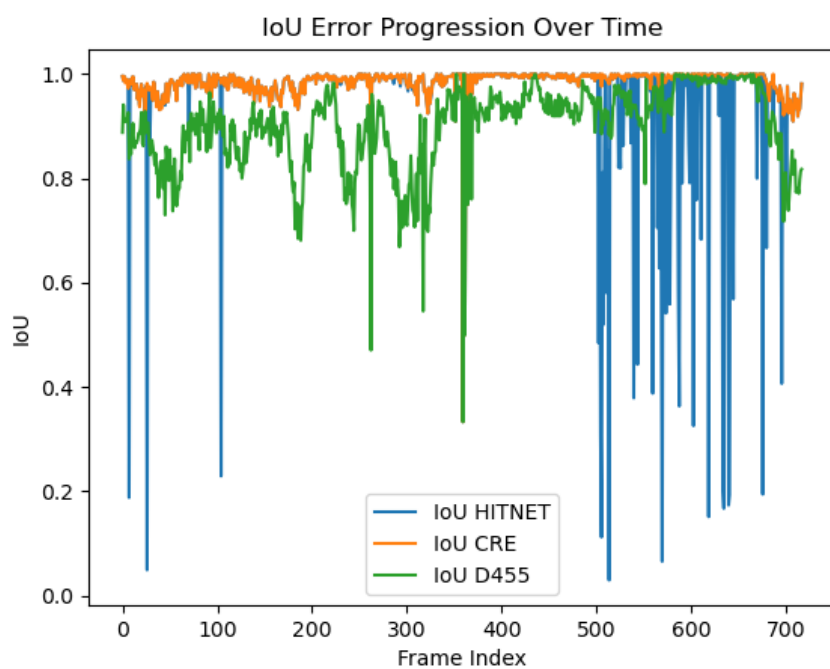


Figure A.8: IoU Error Progression Over Time for Outdoor Data

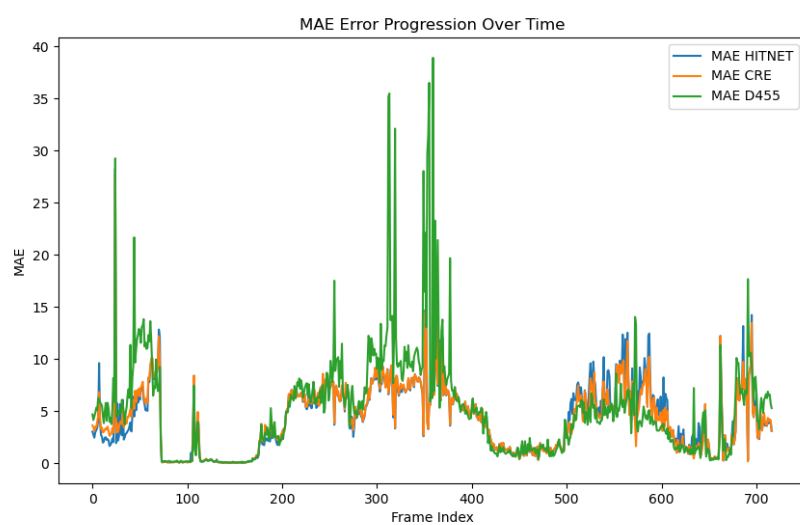


Figure A.9: MAE Error Progression Over Time for Outdoor Data

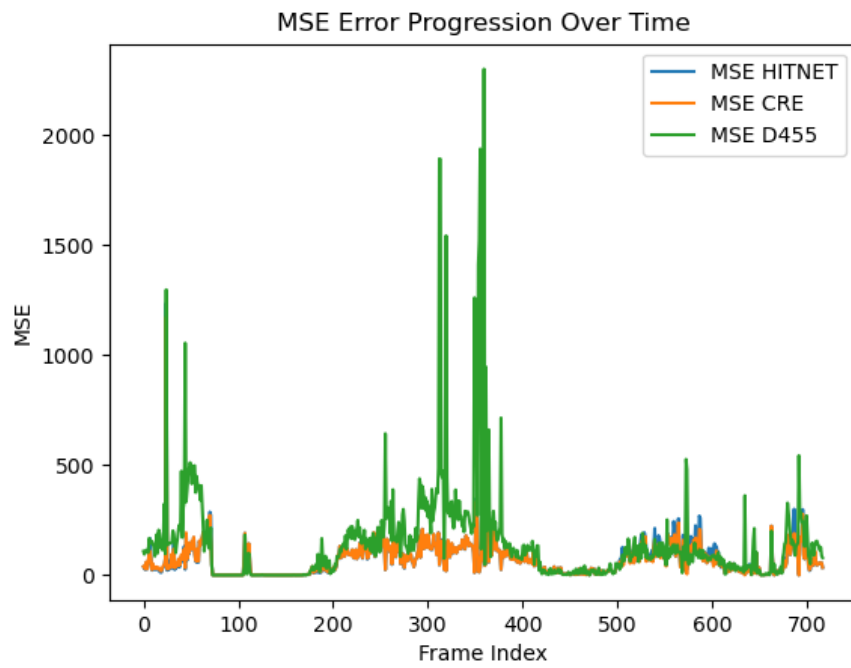


Figure A.10: MSE Error Progression Over Time for Outdoor Data

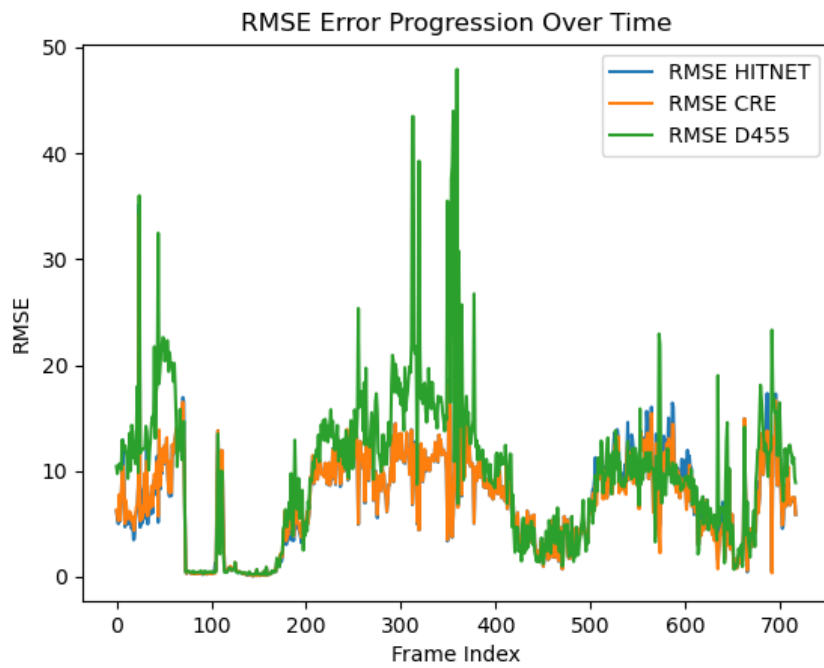


Figure A.11: RMSE Error Progression Over Time for Outdoor Data

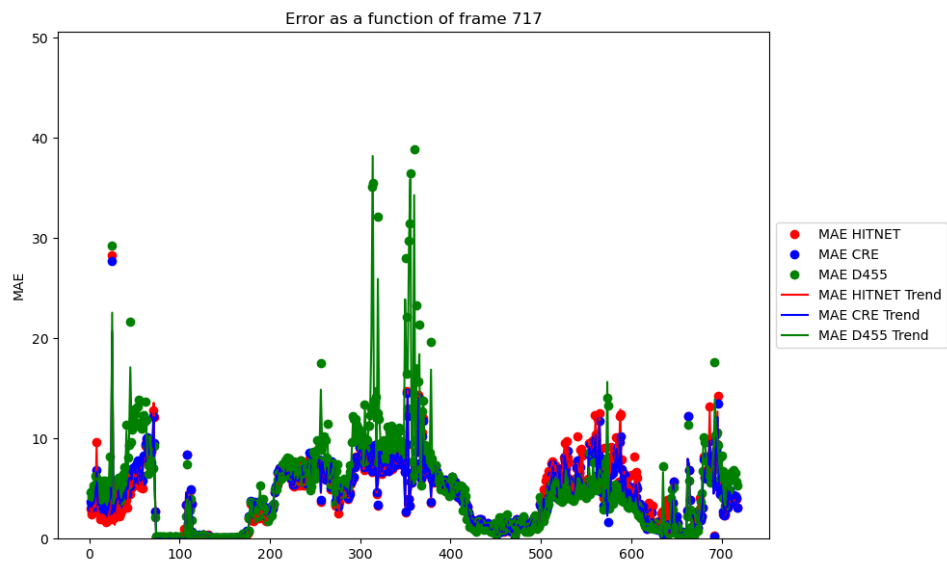


Figure A.12: Error Comparison for All Frames for Outdoor Data