Monocular Depth Estimation Report

Team Name: KERAS
Sai Sudarshan Rao
Tokachichu Sai Chandra Raju
Kotha Aditya
K R Eashwar Sai
Srinidhi

1. Problem Statement

Monocular Depth Estimation (MDE) is a key challenge in computer vision that aims to predict the depth of a scene from a single RGB image. Accurate depth estimation is critical for applications such as autonomous vehicles, robotics, and augmented reality. While significant progress has been made in monocular depth estimation, challenges remain in areas such as depth discontinuities, thin structure handling, and generalization to diverse environments.

In this project, we aim to develop an improved depth estimation model using the **NYUv2** dataset, focusing on enhancing key metrics such as *F-Score*, *MAE*, and *RMSE*, particularly in complex scenes with depth variations.

2. Team Contributions

- **Sai Sudarshan Rao:** Model devlopment and integration of zeodepth.
- **K R Eashwar Sai:** Responsible for dataset preparation, including downloading and preprocessing the **NYUv2** dataset, and ensuring the data pipeline was properly configured.
- **Kotha Aditya:** Model devlopment ,training implementation and helped in deployment.
- **Srinidhi** Handled inference adn evaluation part, enhanced the inference code
- **Sai Chandra Raju** Managed the deployment of the model and model training.

3. Methodology

The solution builds upon the Monodepth Benchmark base code, integrating several key improvements to enhance depth prediction performance.

Network Architecture

We utilize a transformer-style encoder-decoder framework, specifically the Vision Transformer (ViT-L) encoder, to extract rich image features. The **ZoeDepth** model is employed, which divides depth into discrete bins to improve both depth accuracy and generalization across diverse environments.

Supervised Learning with Ground-Truth Depth

Our approach uses *supervised training* with ground-truth depth data from the **NYUv2** dataset. The **SILog loss function** is applied to minimize depth prediction errors, especially for large-scale depth variations common in indoor scenes.

Training Strategy

We adapt the training strategy specifically for indoor scenes using the **NYUv2** dataset:

- Image size: *392×518*
- Batch size: *4*
- Maximum depth: *10 meters*

The learning rate is tailored to ensure efficient convergence without overfitting.

Model Refinements

To handle depth discontinuities and improve predictions for thin structures, we integrate edge-preserving techniques such as **Gradient Loss**. This helps maintain depth detail, especially for objects with sharp edges or narrow structures.

Data Augmentation and Optimization

We employ a comprehensive data augmentation strategy, including random flips, geometric transformations, and color adjustments, to enhance the model's robustness. Additionally, *test-time augmentations* are used to ensure the model generalizes well across different real-world environments.

Evaluation Metrics

The model's performance is evaluated using the following metrics:

- *F-Score* for pointcloud reconstruction.
- *MAE*, *RMSE*, and *AbsRel* for depth accuracy.
- Special attention to *depth discontinuities* and *thin structure accuracy.*

4. Results

The final performance metrics on the **NYUv2** test dataset are as follows:

User	Team Name	F-Score (↑)	MAE (↓)	RMSE (↓)
Aditya	KERAS	21.3	4.1	10.9

Visual examples of model performance:

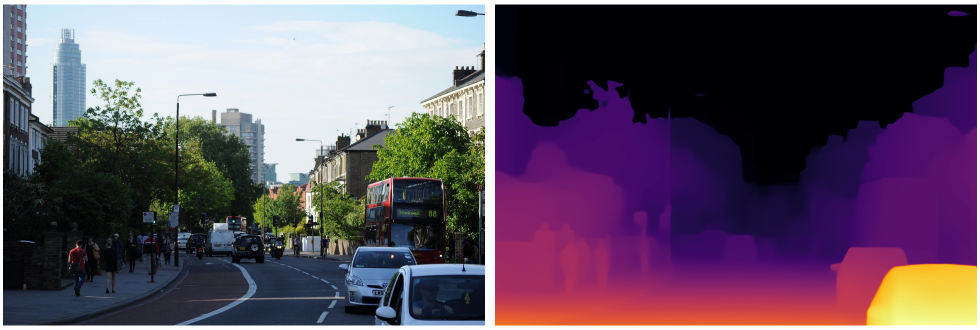

Figure 1: Input Image (Left) and Predicted Depth Map (Right).

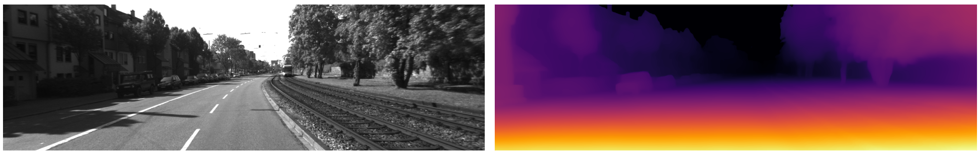

Figure 2: Input Image (Left) and Predicted Depth Map (Right).

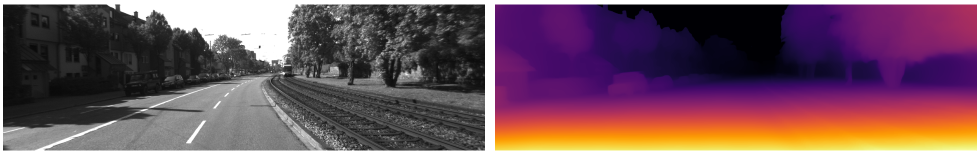

Figure 3: Input Image (Left) and Predicted Depth Map (Right).

5. Conclusion

Team Keras has successfully developed a monocular depth estimation model that improves performance in handling depth discontinuities and thin structures, using the **NYUv2** dataset. The model also generalizes well across various indoor environments. The deployment of the model as a web application further demonstrates its potential for real-time, real-world applications in areas such as robotics and autonomous systems.