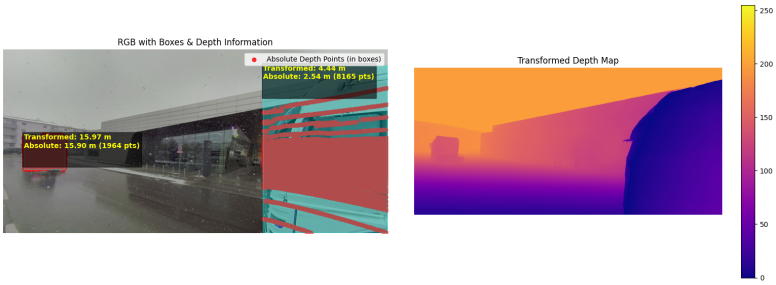
ModelDepthMAN: Comparative Evaluation of ScaleBiasModels for Depth Estimation converting relative depth in absolute depth using MAN TruckScenes Mini

Link for the code and all results: https://github.com/Martons00/modelDeth-MAN-TruckScenes
Practical Implications for Autonomous Driving Applications in the End of the Report

Introduction

This study introduces a comprehensive pipeline for evaluating depth estimation and transformation models applied to the MAN TruckScenes mini dataset, with a systematic comparative analysis of three ScaleBiasModel variants characterized by different computational complexities: a Light ScaleBiasModel optimized for computational efficiency, a Moderate ScaleBiasModel offering balanced performance and computational cost, and a Heavy ScaleBiasModel with high-capacity architecture to maximize predictive accuracy. The complete pipeline encompasses the acquisition and preprocessing of the MAN TruckScenes mini dataset with RGB images and LiDAR point clouds, relative depth estimation using the pre-trained Depth-Anything V2 transformer model, extraction of affine transformation parameters through linear and quadratic polynomial regression techniques, definition and training of three scalable ScaleBiasHead neural network configurations with combined loss functions, multi-metric comparative evaluation against polynomial regression baselines (2D and 3D polyfit) under various operating conditions using RMSE, MAE, and depth consistency measures, validation through YOLO-based object detection for granular per-bounding-box error analysis compared to LiDAR ground truth, and detailed per-image visual analysis with depth map overlays and quantitative comparisons between the three ScaleBiasModel variants and polynomial baselines, establishing a rigorous, interpretable, and scenario-aware evaluation framework particularly relevant for real-world autonomous driving applications in commercial transport scenarios. For the Heavy ScaleBiasModel is also tried a other approach, trading it for 100 epochs with different learning rate (0.001-0.0005) to reach better performance.



Technologies

Architectures of the Three ScaleBiasHead Models

The three implemented ScaleBiasHead models offer distinct architectures, each aiming to balance computational efficiency and predictive accuracy in depth estimation. The Light_ScaleBiasHead adopts a minimalist design, utilizing a reduced number of channels (from 4 to 32), 1x1 and 3x3 convolutions and a streamlined fully connected head that directly maps from 32 to 3 output parameters.
The Moderate_ScaleBiasHead provides a more balanced approach, starting with a convolutional block that uses larger kernels (5x5) followed by standard 3x3 convolutions, and processes channels from 4 to 32 through two max pooling stages. This model concludes with an AdaptiveAvgPool2d that reduces feature maps to a fixed size (4x4), and a fully connected head with 128 hidden neurons and 0.5 dropout for regularization.
The Heavy_ScaleBiasHead is the most advanced architecture, incorporating a MultiScaleFeatureFusion module with four parallel branches (1x1, 3x3, 5x5 convolutions, max pooling followed by 1x1

convolution) to capture features at different spatial scales and a two-stage fully connected head (64→32→3) with dropout, maximizing the model's representational capacity.

DepthAnything2

DepthAnything2 Small is a lightweight yet powerful model for monocular depth estimation, meaning it predicts depth information from a single image. The model is trained on a massive dataset of 595,000 synthetic labeled images and over 62 million real unlabeled images, which allows it to achieve highly accurate and robust depth predictions across diverse scenes. Compared to previous versions and Stable Diffusion-based models, DepthAnything2 Small is significantly faster (about 10 times) and more efficient, while still capturing fine-grained details in depth maps. With only 24.8 million parameters, it is optimized for speed and resource efficiency, making it suitable for real-time applications.

Yolo11n

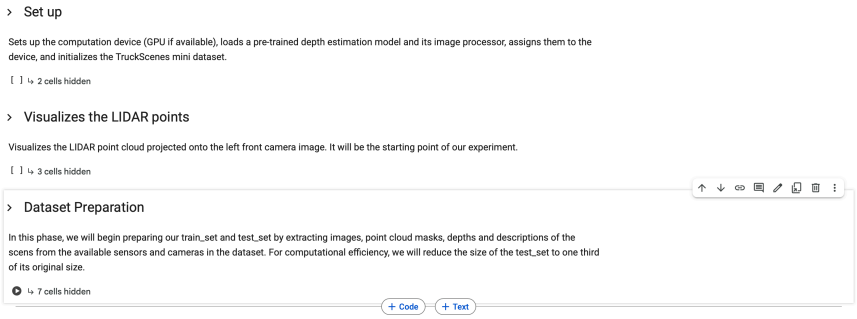
YOLO11n is the nano-sized, ultra-lightweight variant of the YOLO11 object detection model, designed for high efficiency and real-time performance on resource-constrained devices.

MAN TruckScenes Mini

The MAN TruckScenes Mini dataset offers rich multimodal annotations from a comprehensive sensor suite specifically designed for autonomous trucking. Each scene includes synchronized data from 4 RGB cameras, 6 lidar sensors, and 6 radar sensors, providing dense spatial coverage and robust perception in diverse conditions. The lidar sensors deliver high-resolution 3D point clouds, while the radar sensors—featuring 4D capability—capture both range-azimuth and elevation information, yielding around 2,600 points per sample with nearly 360-degree coverage.

Flow of the Script

The implementation prioritizes computational efficiency through strategic memory management and device optimization. Automatic GPU detection enables accelerated processing when available, while explicit memory cleanup prevents accumulation during batch processing. The training pipeline supports variable batch sizes with memory-conscious processing, enabling operation on systems with different computational capabilities. This approach is necessary to address issues related to limited RAM in the Colab environment.



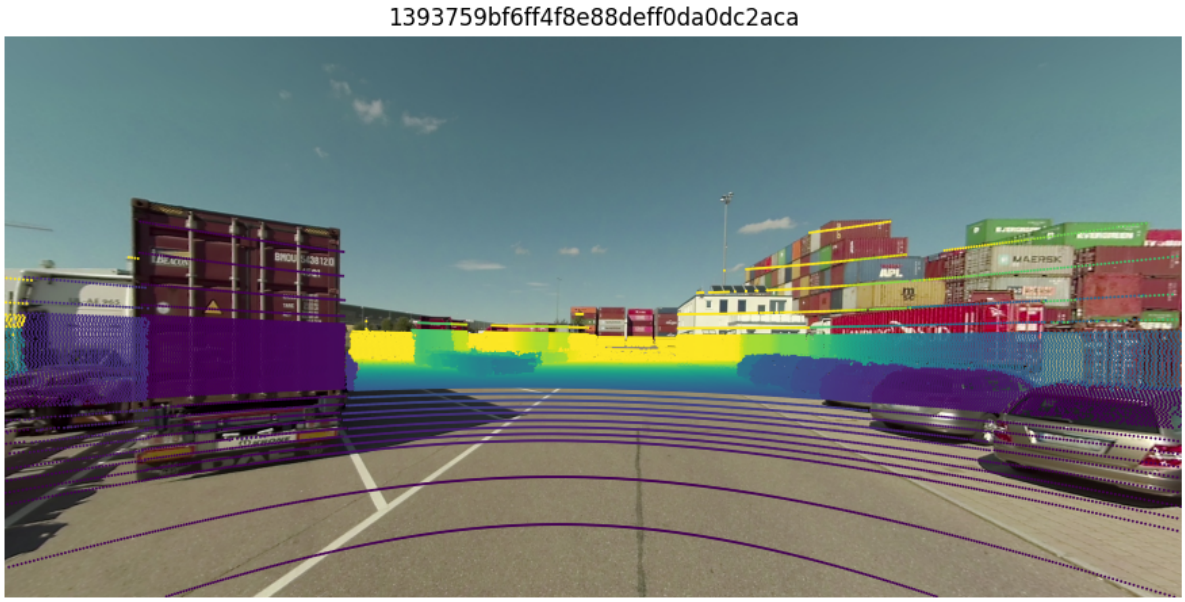
Every parts of the notebook is commented and described.

Dataset Foundation and Preparation

There is strategic train-test split using specific scene selections to ensure balanced representation across different operational conditions. Training scenes (0, 2, 3, 6, 8) encompass diverse environments including terminal areas, city streets, and highways under clear, overcast, and rainy conditions . Test scenes (1, 4, 5, 7, 9) are deliberately chosen to include challenging scenarios such as snow conditions, roadworks, overpass structures, and various lighting conditions including twilight and alternative

illumination. For computational efficiency, the test set is subsampled to 30% of its original size while maintaining representational diversity.

From the dataset using TruckScenes API e some modification for every sample is extracted: t**he image, the mapping point-lidar measurements and the description of the scene.** It's important to notice that for every img there is a limited number of lidar-points that we can use.



Relative Depth Estimation Through Transformer Architecture

During preprocessing, all input images are processed through the Depth-Anything V2 pipeline to generate relative depth heatmaps . These heatmaps capture spatial depth relationships but require transformation to absolute depth measurements for practical applications . The relative depth maps are then inverted using the mathematical operation **depth_rel = depth_rel_max - depth_rel_img** to align with conventional depth representations where larger values correspond to greater distances.

Polynomial Regression for Affine Parameter Extraction

Two distinct polynomial regression are implementes, that approaches to establish baseline relationships between relative and absolute depth measurements using PolyFit. Linear regression models the relationship as **depth_abs = a * depth_rel + b**, providing a simple two-parameter transformation . Quadratic regression extends this to **depth_abs = a * depth_rel^2 + b * depth_abs + c**, capturing non-linear depth relationships through three parameters. The polynomial fitting process involves extracting valid pixel coordinates from LiDAR point cloud projections onto the camera image plane, filtering out invalid depth values, and computing regression coefficients that minimize the mean squared error between predicted and measured absolute depths. These polynomial parameters serve dual purposes: establishing **baseline performance metrics** and providing supervision signals for neural network training.

Neural Network Architecture and Training Strategy

The core of the project lies in the **ScaleBiasHead** neural network, a custom architecture designed to predict affine transformation parameters through learned feature representations. This network processes 4-channel input tensors combining RGB imagery with relative depth information to output three affine

parameters (a, b, c) for quadratic depth transformation. The architectures of 3 models are described in the previous paragraph.

Training employs a hybrid loss function combining depth prediction accuracy with parameter regularization. The **depth loss** component uses mean squared error to measure differences between transformed relative depths and LiDAR ground truth measurements. The **parameter loss** applies Huber loss to encourage predicted affine parameters to remain close to polynomial regression baselines, providing stability and physical plausibility.

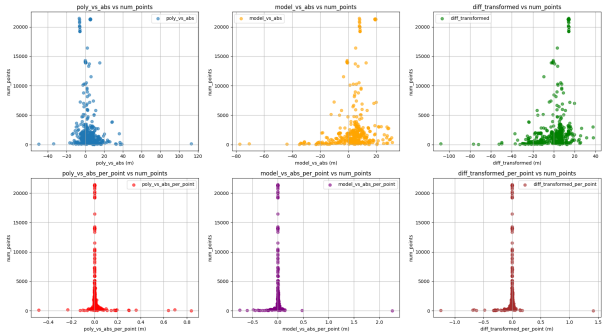
The total loss function is formulated as:
Loss = depth_loss + 0.3 * parameter_loss

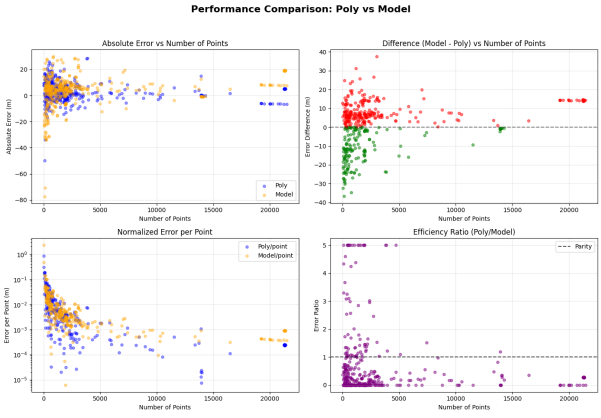
Early stopping with patience=5 prevents overfitting by monitoring validation loss and terminating training when performance plateaus.

Comprehensive Evaluation Methodology

The evaluation framework implements a three-tier approach progressing from abstract loss metrics to physically interpretable measurements. Global evaluation computes test losses across **the entire dataset**, comparing neural network predictions against polynomial baselines using standardized loss functions. **Condition-specific evaluation** filters the dataset by lighting conditions (illuminated, dark, other_lighting) to assess performance under varying environmental scenarios. Then YOLO11n object detection generates bounding boxes for detected objects, enabling per-object depth analysis . For each bounding box, the system computes depth statistics using both model-predicted and polynomial-derived affine parameters, comparing these against LiDAR ground truth measurements within the box boundaries. **Key evaluation metrics** include mean absolute error measured in meters, per-point error normalized by LiDAR point density, and comparative analysis between different approaches. This progression from dataset-wide metrics to object-specific measurements provides comprehensive insight into model performance across different scales and conditions.

Scatter evaluation for all the dataset





Scatter evaluation for polyfit vs Model

Plotting

For each model is taken a random sampling to make a visual evaluation.

Analysis

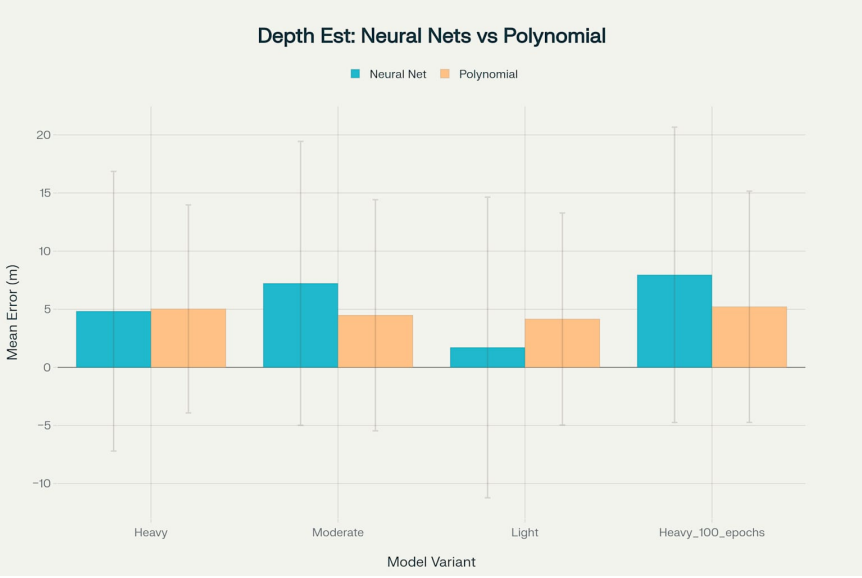
Comparative Analysis

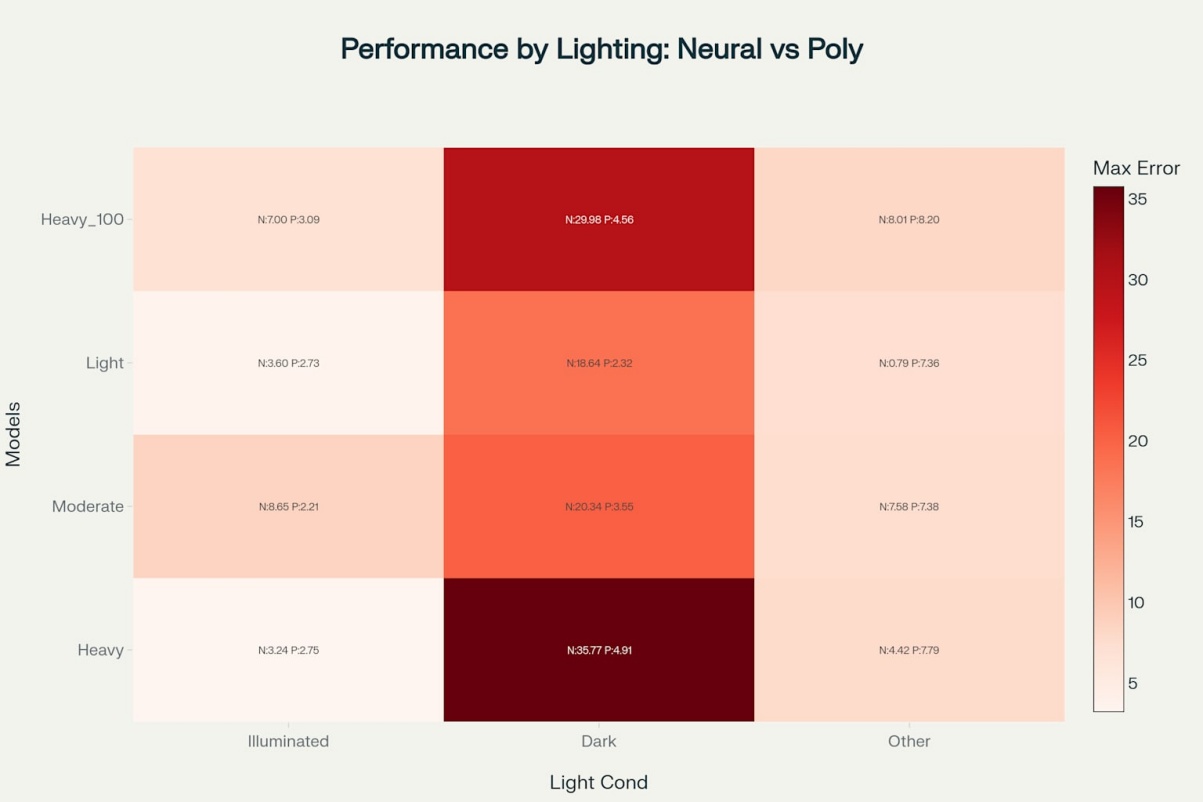
The Light model stands out as a particularly effective solution for real-time applications, achieving the best overall mean absolute error of 1.711m among all neural configurations. Its architecture, optimized for computational efficiency, does not compromise performance; in fact, it excels in mixed lighting conditions with errors as low as 0.787m, outperforming polynomial regression by 9.35x. The model also demonstrates superior capabilities with sparse datasets, achieving errors of 3.791m compared to 10.223m for classical approaches, highlighting the effectiveness of learned neural representations in compensating for limited geometric data density.

The Moderate model doesn't archive good performance. The model is particularly effective with small datasets, achieving errors of 5.961m versus 9.530m for polynomial regression—a 60% improvement. In mixed lighting conditions, the Moderate model remains competitive with nearly equivalent performance to classical approaches.

The Heavy model is the most sophisticated configuration among the neural variants, achieving an almost perfect balance with classical approaches, outperforming them (POLYFIT) in 50.2% of cases. The high-capacity architecture particularly excels in well-lit conditions with errors of just 3.24m, demonstrating superior ability to process high-quality visual information. The Heavy model achieves outstanding results with small and medium datasets, delivering improvements of 67.9% and 51.4% respectively over polynomial regression, showcasing the power of neural feature extraction techniques. In mixed lighting, the model outperforms classical approaches in 63.7% of cases, demonstrating adaptive capabilities that effectively leverage architectural complexity to handle challenging environmental scenarios.

The Heavy model with extended training to 100 epochs maintains distinctive strengths in specific domains, showing particular excellence with sparse datasets where it surpasses polynomial regression by 35.1%. Despite prolonged training, the model remains competitive in mixed lighting conditions, achieving nearly equivalent performance to classical approaches with errors of 8.014m versus 8.204m. But in the overall the performance and the effort the results obtained are not

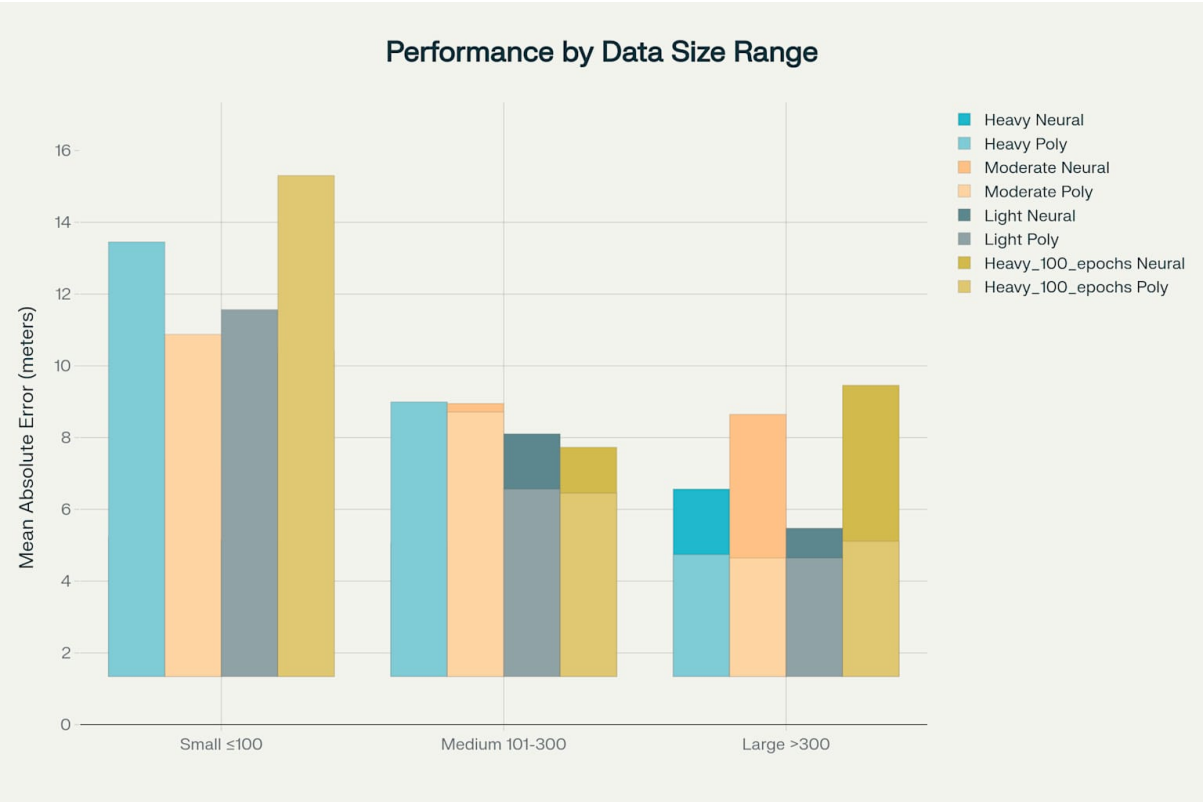




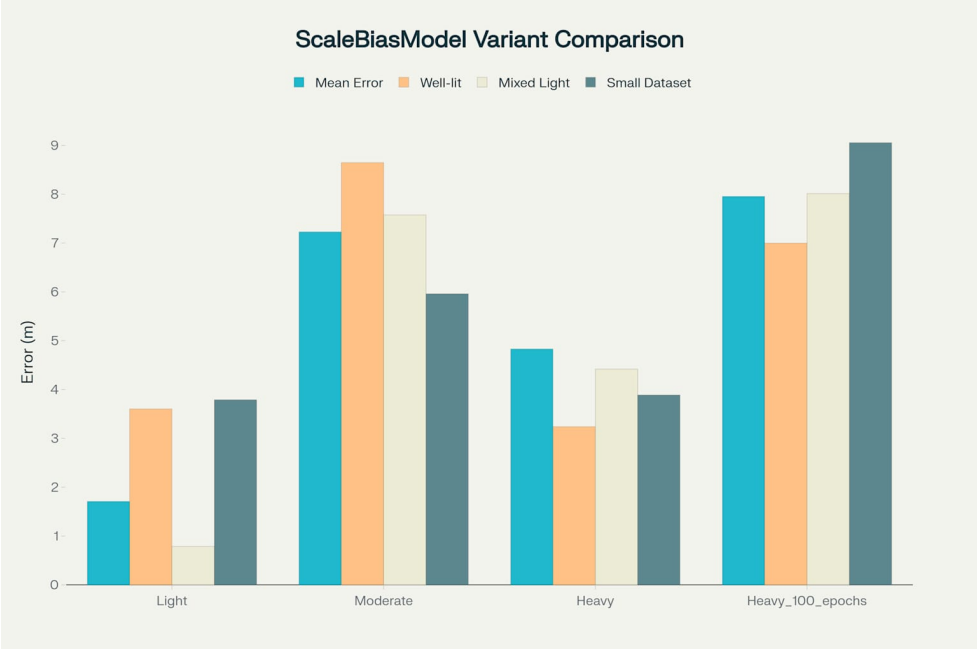
interesting.

Overall Performance Overview

The overall performance comparison between the ScaleBiasModel neural network variants and polynomial regression approaches highlights distinct strengths and trade-offs for each method. The Light model stands out as the most accurate among neural networks, achieving a mean absolute error of



1.711m, while the Heavy_100_epochs model records the highest error at 7.956m, indicating that extended training does not necessarily yield better results. Polynomial regression offers remarkable consistency, with mean errors between 4.157m and 5.210m across all scenarios, and consistently achieves the lowest test loss values, underscoring its stability and reliability. Neural networks, on the other hand, demonstrate a higher sensitivity to architecture and training parameters, resulting in greater variability in performance. Notably, neural models excel with small datasets—where, for example, the Heavy model achieves a 3.893m error versus 12.113m for polynomial regression—but their performance declines with large datasets, where polynomial regression outperforms them. In low-light conditions, all neural models experience significant degradation, while polynomial regression maintains stable results. Conversely, in mixed and optimal lighting, the Light model demonstrates excellent generalization, achieving very low errors and confirming its suitability for real-time applications that require both responsiveness and adaptability.



Practical Implications for Autonomous Driving Applications

The Light model's achievement of 1.711m mean absolute error represents a promising foundation for real-time depth estimation in commercial trucking applications, where computational efficiency and responsiveness are critical for safety-critical decision making . While the neural networks show sensitivity to environmental conditions, particularly in low-light scenarios, these limitations must be contextualized within the experimental constraints of the study, including the relatively small MAN TruckScenes mini dataset, sparse LiDAR point sampling that provides limited ground truth coverage, and computational restrictions imposed by the Colab environment that prevented extensive hyperparameter optimization . The superior performance of neural models with sparse datasets (where the Heavy model achieves 67.9% improvement over polynomial regression) is particularly relevant for autonomous trucking, where sensor occlusion, weather interference, or partial sensor failures may limit data availability . Furthermore, the models' varying performance across different point cloud densities suggests that with optimized sensor fusion strategies, expanded training datasets, and dedicated computational resources for comprehensive parameter tuning, these architectures could achieve substantially improved robustness for deployment in autonomous vehicle perception systems.