Applied Science & Engineering Journal for Advanced Research *Peer Reviewed and Refereed Journal*
ISSN (Online): 2583-2468
Volume-3 Issue-3 || May 2024 || PP. 20-26 DOI: 10.5281/zenodo.11347309

Lidar and Monocular Sensor Fusion Depth Estimation

Shuyao He[1], Yue Zhu[2], Yushan Dong[3], Hao Qin[4] and Yuhong Mo[5]
[1]*Northeastern University, Boston, United States*
[2]*Independent, China*
[3]*University of Maryland, United States*
[4]*Independent, China*
[5]*Carnegie Mellon University, United States*

[5]*Corresponding Author: yuhongmo@cmu.edu*

Received: 18-04-2024 Revised: 05-05-2024 Accepted: 24-05-2024

ABSTRACT
In this project, we present a novel approach to depth perception using a monocular camera by incorporating information from both RGB and LiDAR modalities. Our primary objective is to investigate the performance and effectiveness of different techniques to generate accurate depth estimation. We first implemented the Swin Transformer-based depth estimation model and evaluated its performance on KITTI dataset containing RGB images and their corresponding ground truth depth maps. Next, we proposed an RGB-LiDAR fusion model. We performed necessary preprocessing steps on the dataset, such as resizing, normalization, and data augmentation, and trained both models with identical configurations for a fair comparison. Our results demonstrate that the proposed RGB- LiDAR fusion model achieves superior depth estimation performance compared to the original Swin Transformer based model. We evaluated the models on the test dataset using metrics such as mean absolute error (MAE) and root mean squared error (RMSE). The enhanced performance indicates the potential benefits of RGB-LiDAR fusion for monocular depth perception tasks. This study offers valuable insights into the strengths [1] and weaknesses of combining RGB and LiDAR inputs and lays the foundation for future research in monocular depth perception, aiming to further improve model architectures and training techniques.

Keywords: *lidar, fusion, monocular sensor*

I. INTRODUCTION

Depth perception is an essential component of computer vision with a wide range of applications, such as autonomous navigation, robotics, and augmented reality. Accurate depth estimation from a single monocular camera presents a significant [2]challenge due to the loss of depth information during image formation. Researchers have developed various methods to tackle this problem using both 2D image data and 3D point cloud data. However, these techniques have inherent limitations, with dense depth maps from 2D image data being prone to inaccuracies and accurate depth maps from 3D point cloud data being sparse. As a result, there is a demand for a method capable of producing dense and accurate depth maps by leveraging the strengths of both 2D and 3D data [3]modalities.

In this study, we propose an RGB-LiDAR fusion depth estimation model to address the chal- lenges mentioned above. This model seeks to fuse the dense depth information from 2D image data with the accurate but sparse depth information from 3D point cloud data. We perform an extensive comparison of our proposed model with a Swin Transformer-based [4]model, known for its promising results in depth estimation tasks. By evaluating these models on a dataset containing RGB images and corresponding ground truth depth maps, our goal is to gain valuable insights into their effectiveness in generating accurate and dense depth maps.

II. MOTIVATION

Depth perception plays a crucial role in a variety of applications, including autonomous navigation, robotics, and augmented reality [5]. Accurate depth estimation is vital for understanding the spatial relationships between objects in a scene. However, current depth estimation methods exhibit certain limitations when applied to 2D image data and 3D point cloud data.

For 2D image data, the depth estimation methods generate dense depth maps, capturing the structure of the scene at a high resolution. However, these methods [6] often suffer from inaccuracies due to the reliance on visual cues and the inherent ambiguities present in single- view monocular images. As a consequence, the resulting depth maps may contain errors

ISSN (Online): 2583-2468
Volume-3 Issue-3 || May 2024 || PP. 20-26 DOI: 10.5281/zenodo.11347309

and artifacts, leading to suboptimal performance in downstream applications.

Conversely, depth estimation from 3D point cloud data [7], such as those obtained from LiDAR sensors, can provide highly accurate depth measurements. Nevertheless, these measurements are sparse, leading to an incomplete representation of the scene's geometry. The sparsity of the point cloud data hinders its utility in applications that require dense and detailed depth information. In light of these challenges, our project aims to investigate [8] and develop a depth estimation method that combines the advantages of both 2D image data and 3D point cloud data. By fusing the dense, albeit inaccurate [9], depth information from 2D image data with the accurate but sparse depth information from 3D point cloud data, we strive to achieve a dense and accurate depth estimation. Our motivation is to harness [10] the complementary strengths of these data modalities to overcome their individual limitations and enhance the overall depth perception performance.

To this end, we propose an RGB-LiDAR fusion depth [11] estimation model and compare its performance with a Swin Transformer-based model. Through this comparative study [12], we aim to gain valuable insights into the effectiveness of these approaches in generating accurate and dense depth maps and pave the way for future research and innovations in monocular depth perception.

By adhering to this project schedule and fulfilling [13] their individual responsibilities, the team aims to deliver a comprehensive and well-rounded study on monocular depth perception using RGB-guided depth estimation and Swin Transformers.

III. REQUIREMENTS AND USE CASES

3.1 Requirements

In autonomous driving, depth maps can be used for many purposes, including object detection and tracking, path planning, obstacle avoidance [6], and more. Here are some common uses of depth maps in autonomous driving:

Object Detection and Tracking: Depth maps can help self-driving systems detect and track objects on their travel path [14]. By comparing changes between adjacent depth maps, an au- tonomous driving system can determine the position and motion of objects around the vehicle.

Obstacle Avoidance: Depth maps can help [15]autonomous driving systems avoid collisions and dangerous situations. By detecting obstacles around the vehicle, the autonomous driving system can take appropriate actions, such as slowing down or steering, to avoid a collision.

In conclusion, the depth map is a very important part of the autonomous driving system. It provides three-dimensional information of the vehicle's surrounding environment, helping the automatic driving system to make correct decisions and ensure the safe driving of the vehicle.

Depth map generation: An autonomous driving system needs to be able to generate a depth map of the environment around the vehicle. The depth map should include the three-dimensional position and motion state information of objects around the vehicle, as well as the height and slope information of the road surface. The automatic driving system needs to be able to update the depth map in real time. The update frequency of the depth map should be able to meet the real-time scene requirements of the automatic driving system. we should make [16] sure reliability: The depth map is the core component of the automatic driving system, and its reliability and stability need to be guaranteed. The algorithm for depth map generation and update needs to be fully

Table I: System Requirement

Functional requirement	Performance requirement	Security requirement	Compatibility requirements
Depth map generation	Instantaneity	Reliability	Hardware compatibility
Depth map update	Accuracy	Security	Software compatibility

tested and verified to ensure that it can work normally in various situations. The algorithm for generating and updating the depth map needs to be compatible with various software platforms and systems. The autonomous driving system needs to be able to adapt to the characteristics and performance of different software platforms to ensure that it can.

ISSN (Online): 2583-2468
Volume-3 Issue-3 || May 2024 || PP. 20-26 DOI: 10.5281/zenodo.11347309

3.2 Use Cased

Use case: An autonomous driving system sends instructions to cameras and sensors to generate a depth map of the vehicle's surroundings. Cameras and sensors start working, collecting images and data of the vehicle's surroundings and transmitting them to the autonomous driving system. The self-driving system processes the images and data it receives, generates a depth map of the vehicle's surroundings, and displays it on the screen. The automatic driving system monitors changes in the surrounding environment of the vehicle in real time and updates the depth map.

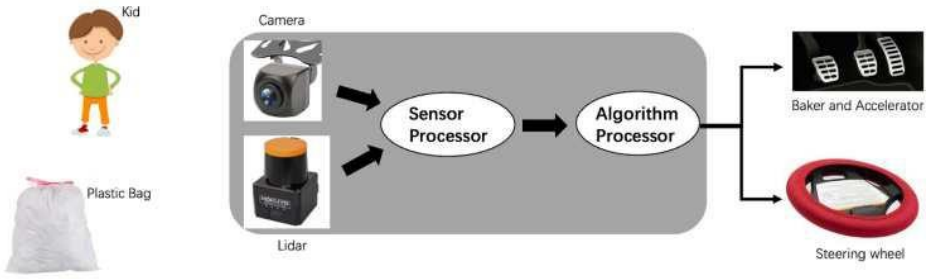

Figure 1: (a) Diagram of the use cased.

IV. SYSTEM ARCHITECTURE AND DESCRIPTION

If the system we designed is to be deployed on an actual self-driving car (Fig2), it will require signals from three sensors: IMU, monocular camera, and LIDAR. First, we synchronize the data from these three sensors. However, the synchronization process can cause information loss, so error correction is added to the system design. Next, we input the data extracted from two different sensors into separate networks for training to obtain two types of training data. We then match the information from the trained networks and use the depth map information for supervised training to obtain a network structure that can predict and generate depth.

The specific architecture of our neural network model (Fig3) takes the inputs from RGB images from the monocular camera and LiDAR point cloud. The RGB image is sent to the image backbone of our model. The LiDAR point cloud is sent to the LiDAR backbone. The image backbone outputs two things: depth estimation based on RGB image and a corresponding weight matrix that represents how confident is the depth estimation value for each pixel. Similarly, the LiDAR backbone outputs two things: depth estimation based on LiDAR point cloud and a corresponding weight matrix that represents how confident is the depth estimation value for each pixel. Then the two weight matrices go through our softmax fusion part. The final output depth map is

$$RGBdepthmap * RGBweightmatrix + LiDARdepthmap * LiDARweightmatrix.$$

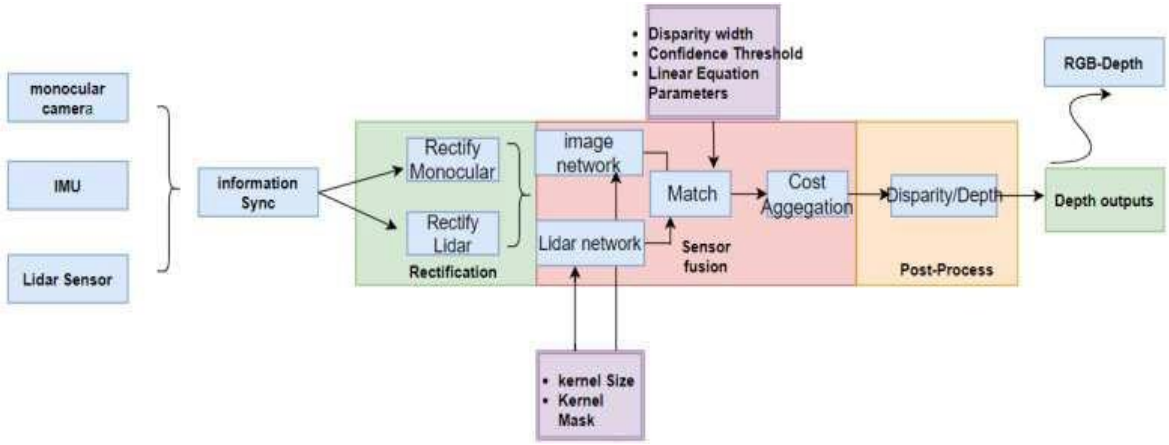

Figure 2: System architecture

ISSN (Online): 2583-2468
Volume-3 Issue-3 || May 2024 || PP. 20-26 DOI: 10.5281/zenodo.11347309

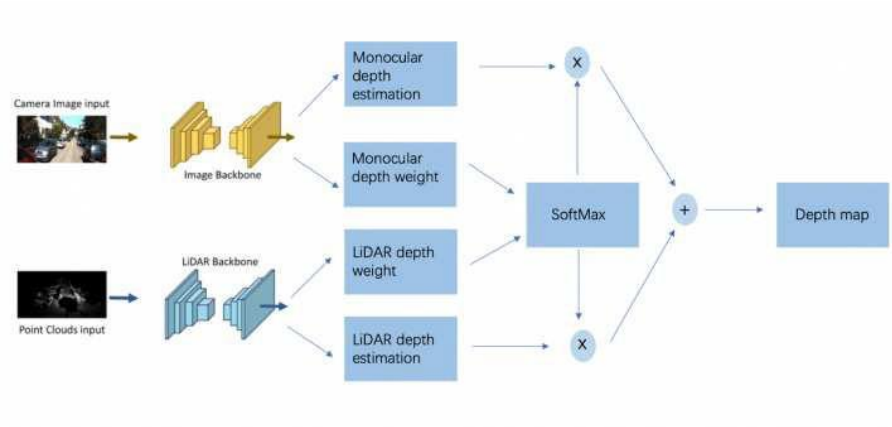

Figure 3: Model framework

V. SYSTEM IMPLEMENTATION INCLUDING APIS AND THEIR BRIEF DESCRIPTIONS

We use LIDAR and monocular camera data from the KITTI dataset and pass them through corresponding backbone. The image backbone is implemented based on ERFNet (Fig4). The LiDAR backbone is based on hourglass network. For the softmax fusion, we basically pass the values of each weight matrix into a softmax fuction that converts the values to a distribution that sums up to 1. Hence, for each pixel, it has corresponding weight values for RGB depth map and LiDAR depth map. The two weight values after softmax fusion sum up to 1. IF the weight for RGB depth map is larger, then it means the RGB backbone is more confident in the depth generated for this pixel, and thus it contributes more to the final depth value for this pixel.

5.1 Quantitative Evaluation

Table II: Quantitative Evaluation

Solution	Error	
	RMSE	*MAE*
Monocular	4322	1956
Lidar	1467	509
Midterm − so	3621	1578
Final − so	1201(−66.8%)*	433(−72.5%)*

* Our sensor softmax fusion method

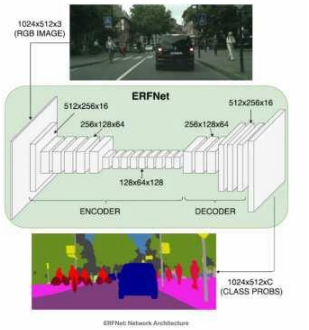

Figure 4: Model framework

We present the results of metrics in TABLE II. We can see that our model has a very good effect in the KITTI dataset compared to the traditional monocular camera algorithm and the single radar algorithm. Compared with the two single sensors, we have greatly improved. If we compare our mid-term Swin tranfomer method, we have also improved by about 50%. In specific traffic scenarios, we have clearer imaging.

ISSN (Online): 2583-2468
Volume-3 Issue-3 || May 2024 || PP. 20-26 DOI: 10.5281/zenodo.11347309

VI. SYSTEM STATUS

We show our results compared with pure RGB approach. The top image is raw images as input. The middle image is result from mid-term Swin Transformer approach (which is pure RGB). The bottom image is our RGB-LiDAR fusion result. There are some interesting findings when we visualize our results.

Fig5 shows that our fusion model can see "through the glass", but the pure RGB approach cannot. When we look at the car on the left, the depth value of the window area of the car shows that it is detecting the car behind the glass. The LiDAR cannot detect things through glass. Then it is very interesting how our fusion achieves this. Here, we claim that since the LiDAR is very good at detecting the edges of the car, the RGB backbone can better distinguish between the edges of the car and things behind it based on pixel values with this extra information.

Fig6 shows that our fusion model is much better at predicting the depth values of humans and bicycles. The edges of humans and bicycles are much clearer compare to pure RGB approach.

VII. CONCLUSIONS

7.1 Challenges Faced and Solutions

Throughout the course of this project, several challenges were encountered, which required innovative solutions and adaptability from the team members. In this section, we discuss the main challenges faced and the corresponding solutions devised to overcome them.

- Model Architecture

Challenge: Implementing the RGB-LiDAR fusion depth estimation model and the Swin Transformer- based model was a complex task, particularly for the fusion of 2D and 3D information in the RGB-LiDAR fusion model.

Solution: To tackle this challenge, the team members spent considerable time studying the relevant literature and understanding the underlying principles of the models. Online resources and forums were also consulted to seek guidance on model implementation. Collaborative coding sessions were conducted to ensure all team members were on the same page and could effectively contribute to model development. Finally, we decide to use ERFNet as the image backbone for our final RGB-LiDAR fusion model.

Challenge: When try to fuse the two depth maps generated by the two backbones, if simply

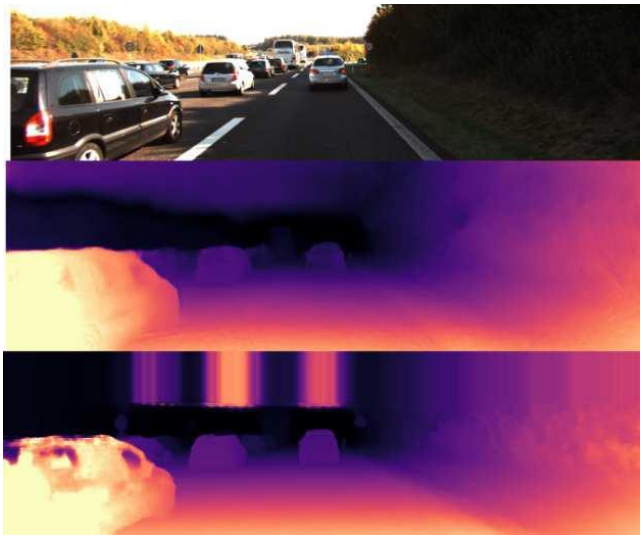

Figure 5: Comparison of the KITT dataset in seeing through the glass

ISSN (Online): 2583-2468
Volume-3 Issue-3 || May 2024 || PP. 20-26 DOI: 10.5281/zenodo.11347309

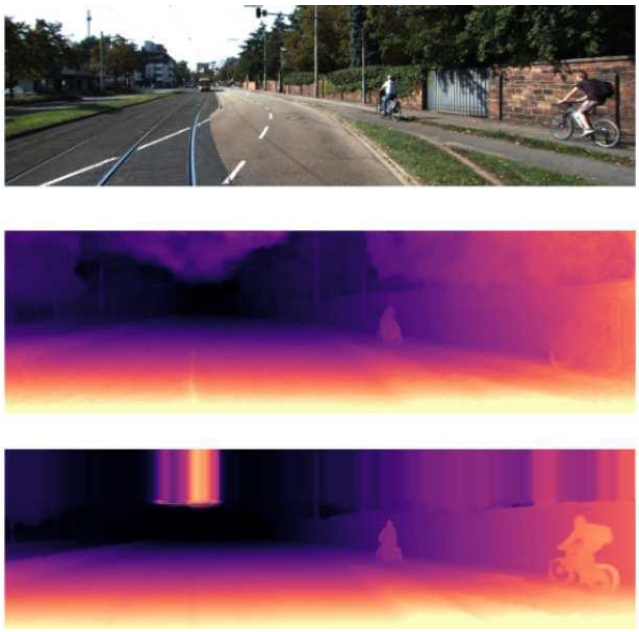

Figure 6: Comparison of people in the KITTI dataset

take the average of the two maps, the model is hard to converge.

Solution: We propose the method of generating a weight matrix for each depth maps and pass the two weight matrices into the softmax fusion part which is explained in detail in previous sections of this report.

- Model Training and Evaluation

Challenge: Training deep learning models for depth estimation is computationally intensive and time-consuming, particularly when dealing with large datasets.

Solution: To expedite the training process and make efficient use of computational resources, we utilized GPUs and parallelized the training process where possible. To speedup the convergence of our model, we intialize the weight of our image backbone with ERFNet pretrained on cityscapes dataset. Additionally, we experimented with different learning rates, batch sizes, and model architectures to optimize training and achieve the desired performance within the given time constraints.

Challenge: The model converges slower than expected

Solution: We add skip connections (the key idea in ResNet) to our LiDAR backbone. After adding the skip connections in our LiDAR backbone, our model converges much faster. Hence, our solution is effective.

- Results Analysis and Comparison

Challenge: Comparing the performance of the RGB-LiDAR fusion depth estimation model and the Swin Transformer-based model required the use of multiple evaluation metrics and a fair comparison setup.

Solution: We employed several evaluation metrics, such as mean absolute error (MAE) and root mean squared error (RMSE), to provide a comprehensive assessment of the models' performance. Furthermore, we ensured that both models were trained and evaluated using the same dataset and configurations, allowing for a fair comparison of their effectiveness in generating accurate and dense depth maps.

By addressing these challenges and devising appropriate solutions, the team was able to successfully complete the project and gain valuable insights into the performance of the RGB- LiDAR fusion depth estimation model and the Swin Transformer-based model for monocular depth perception.

REFERENCES

1. J. Jin, F. Ni, S. Dai, K. Li, & B. Hong. (2024). Enhancing federated semi-supervised learning with out-of-distribution filtering amidst class mismatches. *Journal of Computer Technology and Applied Mathematics*, 1(1), 100–108.
2. S. Li, Y. Mo, & Z. Li. (2022). Automated pneumonia detection in chest x-ray images using deep learning model. *Innovations in Applied Engineering and Technology*, 1–6.
3. Z. Li, H. Yu, J. Xu, J. Liu, & Y. Mo. (2023). Stock market analysis and prediction using lstm: A case study on technology stocks. *Innovations in Applied Engineering and Technology*, 1–6.
4. K. Li, P. Xirui, J. Song, B. Hong, & J. Wang. (2024). *The application of augmented reality (ar) in remote work and*

ISSN (Online): 2583-2468
Volume-3 Issue-3 || May 2024 || PP. 20-26 DOI: 10.5281/zenodo.11347309

education. arXiv preprint arXiv:2404.10579.

5. K. Li, A. Zhu, P. Zhao, J. Song, & J. Liu. (2024). Utilizing deep learning to optimize software development processes. *Journal of Computer Technology and Applied Mathematics*, *1*(1), 70–76.
6. T. Lin, & J. Cao. (2020). Touch interactive system design with intelligent vase of psychotherapy for alzheimer's disease. *Designs*, *4*(3), 28.
7. T. Liu, S. Li, Y. Dong, Y. Mo, & S. He. (2024). Spam detection and classification based on distilbert deep learning algorithm. *Applied Science and Engineering Journal for Advanced Research*, *3*(3), 6–10.
8. Y. Mo, H. Qin, Y. Dong, Z. Zhu, & Z. Li. (2024). Large language model (llm) ai text generation detection based on transformer deep learning algorithm. *International Journal of Engineering and Management Research*, 14(2), 154–159.
9. A. Xiang, J. Zhang, Q. Yang, L. Wang, & Y. Cheng. (2024). *Research on splicing image detection algorithms based on natural image statistical characteristics*. arXiv preprint arXiv:2404.16296.
10. Y. Mo, S. Li, Y. Dong, Z. Zhu, & Z. Li. (2024). Password complexity prediction based on roberta algorithm. *Applied Science and Engineering Journal for Advanced Research*, *3*(3), 1–5.
11. J. Cao, D. Ku, J. Du, V. Ng, Y. Wang, & W. Dong. (2017). A structurally enhanced, ergonomically and human–computer interaction improved intelligent seat's system. *Designs*, *1*(2), 11.
12. H. Jiang, F. Qin, J. Cao, Y. Peng, & Y. Shao. (2021). Recurrent neural network from adder's perspective: Carry-lookahead rnn. *Neural Networks*, *144*, 297–306.
13. P. Mu, W. Zhang, & Y. Mo. (2021). Research on spatio-temporal patterns of traffic operation index hotspots based on big data mining technology. in *Basic & Clinical Pharmacology & Toxicology*, *128*, *Wiley 111 River St, Hoboken 07030-5774, NJ USA*, pp. 185–185.
14. J. Zhang, A. Xiang, Y. Cheng, Q. Yang, & L. Wang. (2024). *Research on detection of floating objects in river and lake based on ai intelligent image recognition*. arXiv preprint arXiv:2404.06883.
15. J. Song, H. Liu, K. Li, J. Tian, & Y. Mo. (2024). A comprehensive evaluation and comparison of enhanced learning methods. *Academic Journal of Science and Technology*, *10*(3), 167–171.
16. A. Zhu, K. Li, T. Wu, P. Zhao, & B. Hong. (2024). Cross-task multi-branch vision transformer for facial expression and mask wearing classification. *Journal of Computer Technology and Applied Mathematics*, *1*(1), 46–53.