Anytime Stereo Image Depth Estimation

Anirudh Iyengar Kaniyar Narayana Iyengar
Robotics and Autonomous Systems (AI)
Arizona State University
akaniyar@asu.edu

Akshara Trichambaram
Robotics and Autonomous Systems (AI)
Arizona State University
atricham@asu.edu

Nivina Christy
Robotics and Autonomous Systems (EE)
Arizona State University
nchris17@asu.edu

Roop Sumanth Gundu
Robotics and Autonomous Systems (AI)
Arizona State University
rgundu2@asu.edu

1 INTRODUCTION

In this report, we introduce a cutting-edge deep-learning model "Any-Net" that addresses the challenge of disparity and depth estimation in stereo images. As a widely used process for autonomous vehicles, this task demands high accuracy and real-time performance. Leveraging the power of deep learning, our model achieves superior disparity and depth estimation without increasing the computational complexity, thereby enabling real-time implementation. This method makes use of both an encoder and decoder for feature extraction and has resulted in improved accuracy compared to the baseline which uses only the encoder section of the U-Net model[1] for their feature extraction, thus showing promising results. We took Kitti 2012 stereo images as our dataset and scene flow as pre-trained checkpoints. Stereo images consist of a pair of images that capture the same scene with a slight change in the perspective of one from the other. This technique mimics the human capability of estimating the depth of an object using the vision in their eyes.

Our work represents a major breakthrough in the field of autonomous driving, particularly in decision-making thus offering a significant potential for the future of autonomous vehicles. By providing a robust and accurate estimation of the scene's depth and disparity, our model can enhance the safety and efficiency of autonomous systems, opening up new opportunities for innovation and progress in this rapidly evolving field. As technology demands innovations in transportation navigation, advanced models like Any-Net will contribute to significant results in the field. Through the scope of this research, we intend to increase the performance of the AnyNet architecture leading to better-fine-tuned results. The additions to the model are such that it does not create a significant change in its computational complexity while yielding better quantitative and qualitative results.

2 RELATED WORK

For over a century, researchers have been striving to make the vision of self-driving vehicles a reality. One of the key challenges in this area is the disparity and depth estimation from stereo images, which is crucial for enabling autonomous systems to perceive their environment and make informed decisions. In recent years, a wealth of research has been conducted in this field, leading to several significant advancements:
Nighttime stereo depth estimation using a joint translation-stereo learning [2] approach addresses the specific challenges posed by driving at night, providing accurate depth estimation even in low-light conditions.
Another notable method, **Pyramid Stereo Matching Network** [3], was the state-of-the-art (SOTA) for a time, leveraging a multi-scale approach to improve disparity and depth estimation accuracy.
Building on the PSMNet, The **Group-wise correlation stereo network** [4] proposes a new architec-

ture that improves the efficiency and accuracy of stereo matching.
Meanwhile, the **Anytime stereo image depth estimation on mobile devices**[1] focuses on enabling real-time depth estimation on mobile devices with limited computational resources. This is achieved by developing a lightweight model that can run on a mobile device without sacrificing accuracy.
The U-Net approach being the backbone of our model is derived from **U-Net: Convolutional Networks for Biomedical Image Segmentation** [5] which is used extensively for a variety of segmentation and disparity estimation problems. The use of ML training techniques in the field of autonomous driving systems has seen a boom in recent years.

3 BASELINE RESULTS

The baseline model for our project is the AnyNet model (Figure 9), which utilizes a novel approach to disparity and depth estimation using convolutional neural networks. The dataset employed to obtain the results through anynet is the KITTI 2012 dataset. The model works by extracting features from the input stereo image pairs using convolutional layers and down-sampling the images. For this function, we leverage the pre-trained SCENEFLOW dataset that acts as the training checkpoint. The down-sampled images are then passed through a warping layer and a disparity network, which produces a disparity map for each image pair. The up-sampled images and residual maps from the disparity network are then combined and aggregated to obtain the final disparity map. The AnyNet model serves as a strong baseline for our project, as it provides an efficient and accurate method for disparity and depth estimation. By leveraging deep learning techniques and combining them with advanced warping and disparity networks, the AnyNet model achieves impressive results on a variety of stereo-images. Figure 8 shows the result obtained from the base model. Figure 1

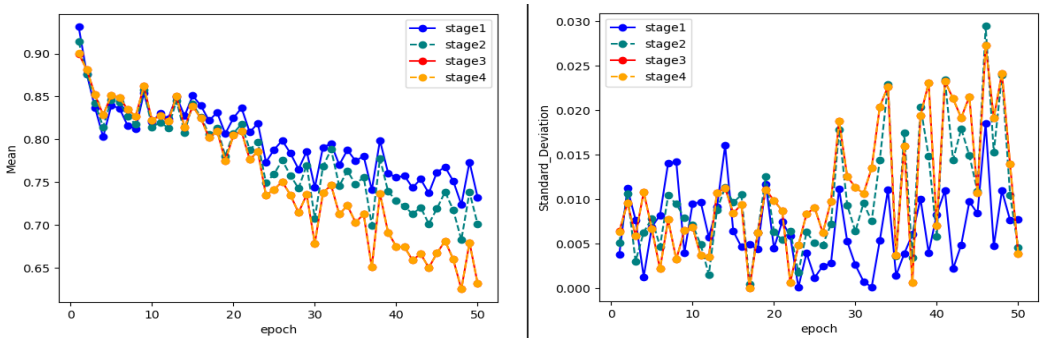

Figure 1: Mean[L], SD[R] - base model

shows the graph showing the mean and SD of the losses. It denotes that with every run, the error function decreases which ensures better results.The means and SD throughout the stages show a divergence before reaching a stable point. Our goal is to build upon the strengths of the AnyNet model and further improve the accuracy and efficiency of disparity and depth estimation for autonomous driving applications. By refining the existing model architecture and incorporating new features and techniques, we aim to enhance the performance of the baseline model and take another step toward the vision of fully autonomous vehicles.

4 OUR APPROACH

4.1 PHASE I

Initially, we tried tuning hyper-parameters on the existing model, without altering baseline architecture to see if we get better results. After several attempts, we found a slightly better qualitative result with the Adagrad optimizer, even though it did not significantly improve the baseline. Figure 11 represents the qualitative result obtained while tuning the model. Our main intention was to improve the edge cases of the result, thus we decided to concentrate more on the feature extraction. Since it is a real-time implementation, we cannot increase the computational complexity and run time, also we need an output at any time as expected as per our base model requirements. Keeping these things in mind we came up with the architectural changes for phase II implementation.

4.2 PHASE II

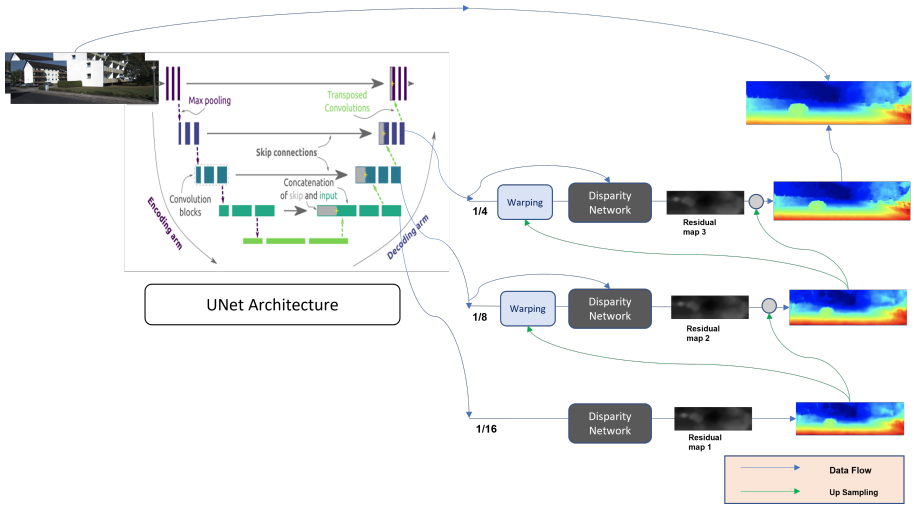

Figure 2: Delta architecture

Here, we have sought to enhance the accuracy and efficiency of disparity and depth estimation for autonomous driving applications by using advanced deep learning techniques. Specifically, we have focused on upgrading the feature extraction stage of the Anynet model to a fully connected U-Net with an encoder and decoder (figure 2). The decoder part of our model consists of four block layers, each of which produces an upsampled image that is passed through a warping layer and a disparity network. The output of each block layer is then combined and aggregated to obtain the final disparity map. To evaluate the effectiveness of our approach, we conducted a series of experiments on a stereo KITTI 2012 dataset. Our results demonstrate a significant improvement in accuracy and efficiency compared to the baseline AnyNet model. On comparing the mean (figure 3) and SDs

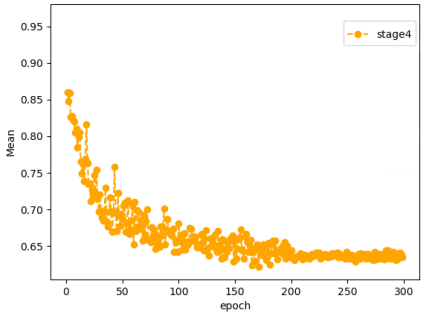

Figure 3: Mean - Our model

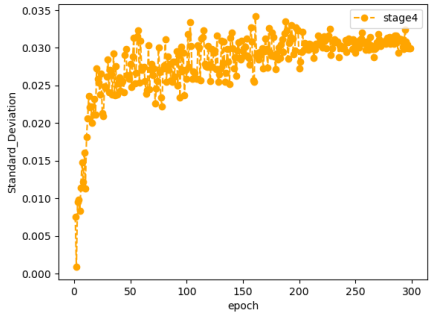

Figure 4: SD - Our model

(figure 4) of the last stage computations, we can see greater stability of the values with the delta implementation. The mean reaches convergence to the optimum value more quicker. The SDs show a peak through the test set showing an increase in the spread of data similar to the baseline model. We also compared the runtime performance of our model to that of the baseline, finding that the baseline approach requires significantly fewer computational resources. Smooth L1 loss function was used in both models and the comparison map is shown in figure 5, it is evident that our model performance did not compromise in terms of efficiency compared to baseline yet obtained accurate depth with better resolution. Overall, our approach represents a significant advancement in the field of disparity and depth estimation for autonomous driving applications. By leveraging advanced deep learning techniques and carefully refining the model architecture, we have achieved a substantial improvement in performance compared to the existing state-of-the-art . We believe that our approach has the potential to greatly benefit the future of autonomous vehicles, enabling them to operate

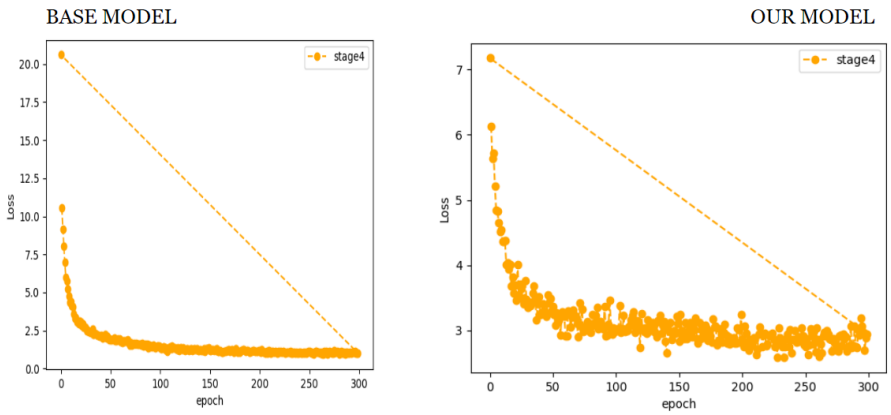

Figure 5: Performance Comparison between Base and Our model

safely and efficiently in a wide range of environments and scenarios. Figure 6 and figure 7 shows the quantitative results obtained for the ground truth. Figure 12 shows the qualitative comparison between the ground truth , base model and our delta implementation for multiple images. In addition to these

Figure 6: Ground truth

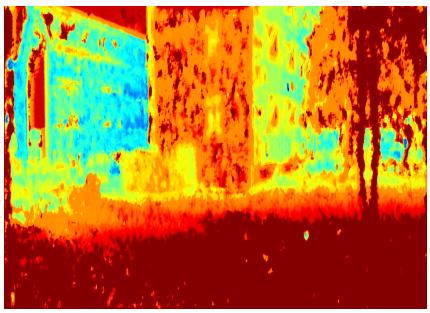

Figure 7: Delta model result

changes, we also explored methodologies such as using Swin Transformer for feature extraction and data augmentation to improve computational accuracy. However, these techniques were found to be computationally expensive and not suitable for implementation in AnyNet for anytime computation.

5 Conclusion and Future work

In conclusion, our deep learning model for disparity and depth estimation has shown promising results in achieving real-time performance on a GPU. Our approach of using a fully connected U-Net with an encoder and decoder for feature extraction along with the fine-tuning of the architecture and hyperparameters of the model has resulted in improved accuracy compared to the baseline.

Moving forward, there is potential for further improvement by exploring different feature extraction techniques such as transformers and using scale-invariant log loss as the loss function. Additionally, we plan to extend our model to work with video data, which will require further optimization to ensure real-time performance.

Overall, our work represents an important step toward the development of more efficient and accurate autonomous driving systems. Latest implementations of UNett++ [6] can be also be a possible improvement to the existing model to enhance the feature extraction module. Additionally, we can explore the use of transfer learning and pre-trained models to further boost the performance of our model on different datasets for the same application.

References

[1] Y. Wang, X. Guo, W. Chen, and Y. Dai, "Anytime stereo image depth estimation on mobile devices," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 708–709, IEEE, 2020.

[2] A. Sharma, M. Gupta, and N. Ahuja, "Nighttime stereo depth estimation using joint translation-stereo learning: Light effects and uninformative regions," *arXiv preprint arXiv:1804.03073*, 2018.

[3] J.-R. Chang and Y.-S. Chen, "Pyramid stereo matching network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5410–5418, 2018.

[4] X. Guo, K. Li, W. An, and Y. Yang, "Group-wise correlation stereo network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 10328–10337, 2019.

[5] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, vol. 9351, pp. 234–241, 2015.

[6] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," *IEEE Transactions on Medical Imaging*, vol. 38, no. 2, pp. 424–435, 2019.

A Appendix

Figure 8: Input image, residual, disparity stages respectively of base model

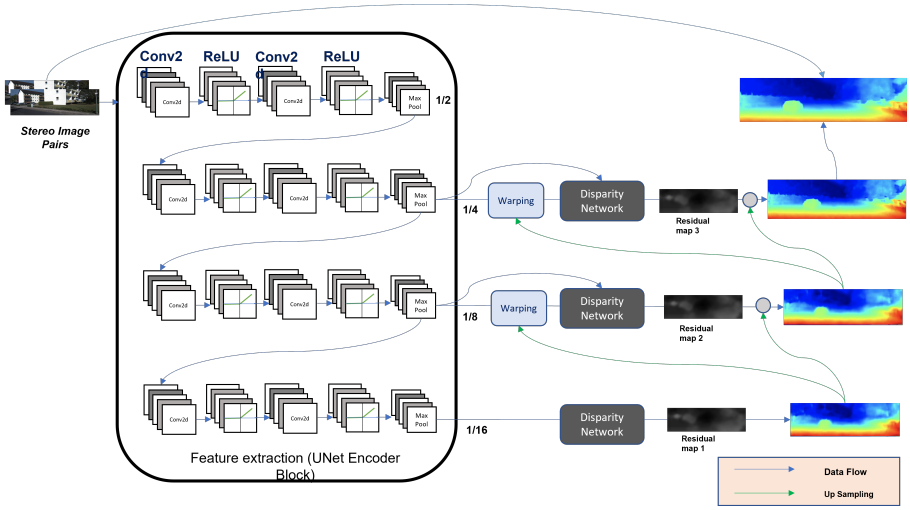

Figure 9: Baseline Model

Figure 10: Ground truth

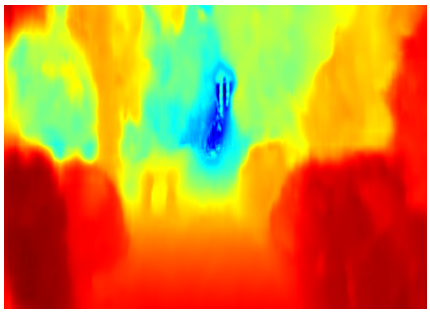

Figure 11: Hypertuned result

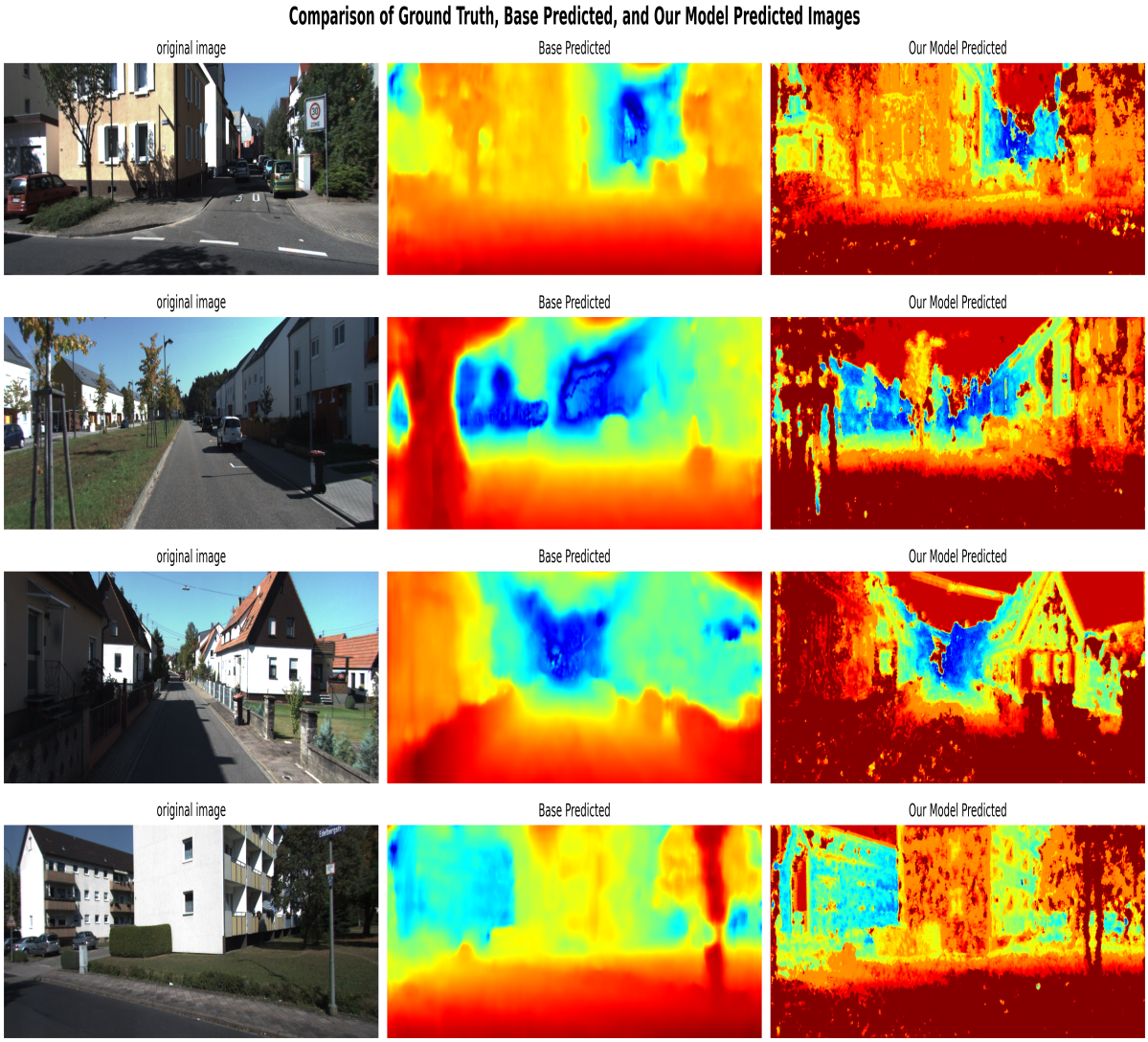

Figure 12: Qualitative comparison: Ground truth, Base model, and Delta implementation