High Quality Depth Estimation using Monocular Camera via "Transfer Learning"

Aureziano Faria de Oliveira
Departmento de Informática
Universidade Federal do Espírito Santo
Vitória, Brasil
aureziano.oliveira@edu.ufes.br

Claudine Badue
Departmento de Informática
Universidade Federal do Espírito Santo
Vitória, Brasil
claudine@lcad.inf.ufes.br

Alberto F. De Souza
Departmento de Informática
Universidade Federal do Espírito Santo
Vitória, Brasil
alberto@lcad.inf.ufes.br

***Abstract*— Depth estimation using monocular images is still a challenge that has gained a better understanding with deep convolutional neural networks. Applying a new transformerbased block architecture called adaptive bins (or Adabins)[1], we try to achieve the foundation of human depth perception from a single image. Extensive datasets need to be used for training, so NYUDepth[2] and KITTI[3] are used to obtain the proposed model weights. We tested the performance of the proposed model by integrating it into the navigation system of our autonomous vehicles and interacting with data collected from the Intelligent Autonomous Robotic Autonomobile (IARA) and the Automic Robotic Truck (ART)[4].**

Keywords— depth estimation,deep learning,convolutional neural networks, adabins

I. INTRODUCTION

Depth perception can have several applications in the scope of computer vision, to detect objects and people in relation to their proximity. Thus, in this context, a solution with great efficiency is the use of deep neural networks to estimate an image depth map. Adabins [1] is a neural network that has a block architecture to estimate the adaptive bin width, having RGB images as input and depth images as output.

II. RELATED WORKS

A. Learning Depth from Single Images with Deep Neural Network Embedding Focal Length

Lei He et all [3] proposes using CNN with a global depth estimate and another with greater accuracy locally using the NYU Depth and KITTI datasets.

B. AdaBins: Depth Estimation using Adaptive Bins

Shariq Farooq Bhat et all [1] It proposes the use of CNN with the Adabins model, where bin widths are calculated according to the input image, using NYU and KITTI as a dataset.

TABLE I. KITTI DATASET PERFORMANCE COMPARISON

Method	δ1↑	δ2↑	δ3↑	REL ↓
Eigen et al. [4]	0.702	0,898	0,967	0,203
Liu et al. [2]	0,680	0,898	0,967	0,201
AdaBins [1]	**0.964**	**0.995**	**0.999**	**0.058**

III. METHODOLOGY

A. Dataset and Annotations

KITTI data[3] is composed of external scenes obtained from cameras placed in cars and with a depth sensor, this set has about 93k of images.

The NYU Depth v2 dataset[2] is composed of 464 scenes, obtained by Microsoft Kinect. Thus, due to the large number of frames obtained, approximately 4k of RGB-D images result, with the treatment and random cuts reaching around 48k.

Thus, these KITTI[3] and NYU[2] Depth data have a reduction in resolution which is 640x480 pixels and 960x224 pixels respectively, to 320x224 in the training phase.

B. Evaluation Metrics

Comparing Adabins[1] results with other methods, we noticed that it is very efficient. As we can see from Table I, Adabins easily outperforms other state-of-art methods.
When used indoors, suitable weights are those generated with the NYU[2] dataset. For our use, the weights generated with the KITTI[3] dataset are the most suitable.

C. Model Architecture

The architecture of the Adabins[1] model used has an encoder-decoder block and a bin width estimator block, the first part being a simple depth regression network. Thus, the differential of the method is that the output is not just an image with depth, but a tensor.

From Figure 1 we have:

- **Mini-ViT[1].** Within the depth range estimates the subintervals.
- **Width of the boxes[1].** Using Mini-ViT a concentration of the network in the interesting sub-intervals is obtained.

Reach attention maps: promotes integration between global information and local information, to build the depth map.

- **Hybrid Regression[1].** its final depth value is calculated for each pixel using Softmax scores from linear combinations.

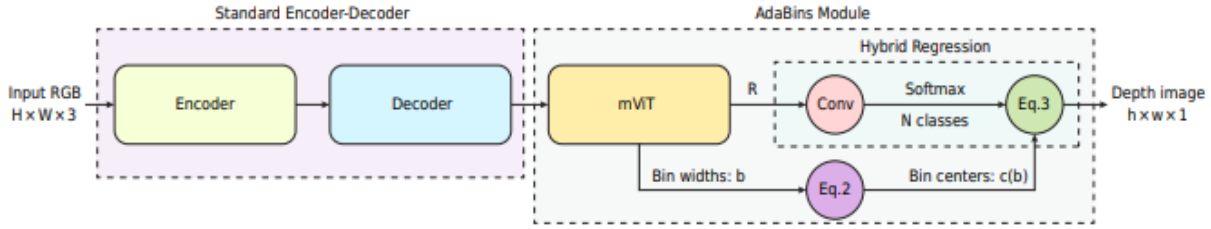

Figure 1- View of the Adabins architecture [1]

IV. EXPERIMENTS AND RESULTS

A. *Integrating with CARMEN.*

We created a module integrated to the CARMEN navigation system used in IARA and ART[4], which loads and consumes the Adabins network implemented in by the author. Our module receives messages from CARMEN cameras in real time and, using the network, delivers depth maps.

Our module's source code and usage can be found et CARMEN_LCAD's github page [https://github.com/LCADUFES/carmen_lcad/tree/master/src/deep_mapper]. The colab notebook and pythons scripts to run detached of CARMEN can be downloaded from here [https://github.com/thiagorjes/DeepMapper]. The results of the integration tests are available at [https://drive.google.com/drive/folders/1Hmu0vR7Jdi5cxkCUBXkUnr6Zr9DRmmWE?us p=sharing].

B. *Depth Map from IARA and ART's images.*

Tests were performed with ART and IARA[4] logs with impressive results.

The Fig. 2 shows the first try, using logs from IARA. The Fig. 3 shows the second try, using logs from ART.From the ART's generated depth maps, a sample was taken and the 3D reconstruction of this sample can be performed, as shown in Figure 4. A video of the 3D reconstructions is available from [https://drive.google.com/drive/folders/1Hmu0vR7Jdi5cxkCUBXkUnr6Zr9DRmmWE?u sp=sharing].

Fig. 2.IARA's RGB and Depth Map images comparison.Blue is closer, Orangeis far.

As we can see, Adabins[1] can be easily integrated and used in carmen with both vehicles (iara and art). This makes it a good choice to create a subsystem that can replace LIDAR in specific situations.

For more details about the tests performed at IARA[4] and ART, watch the videos available on the link [https://drive.google.com/drive/folders/1Hmu0vR7Jdi5cxkCUBXkUnr6Zr9DRmmWE?us p=sharing].

If you have a computer with GPU installed and the correct version of Ubuntu Linux, download the ART logs and test the system in playback mode according to the tutorial available

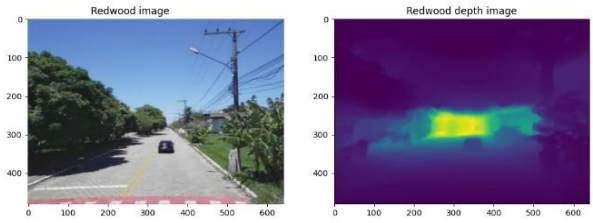

Fig. 3. ART's RGB and Depth Map images comparison. Blue is closer, yellow is far.

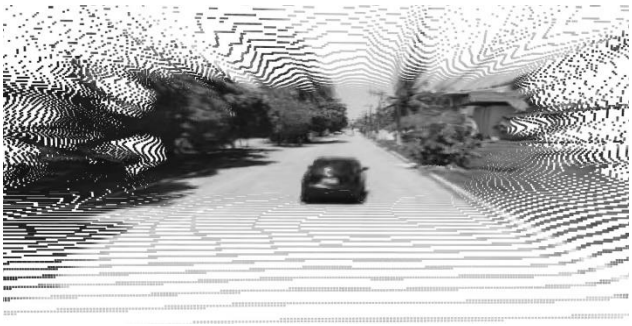

Fig. 4. 3D reconstruction scene from ART's log.

V. CONCLUSION AND FUTURE WORKS

In this article, we present an overview of using Adabins to estimate monocular camera depth maps. Adabins provides a robust solution to the non-stereo image stream depth map estimation problem and can be used in the context of autonomous cars.

We were able to integrate it with the CARMEN autonomous system, used both in IARA and ART, our autonomous vehicles, and even using the weights provided by the author, we were able to reconstruct real scenes from the data collected by the vehicles.

As goals for future work, we will adapt the training process to use the vast database of IARA and ART, thus optimizing the results for our operating environment.

Another possibility is to adjust the network input to an image format that ignores regions where the "ground truth" does not provide a good depth estimate (i.e. sky). We hope with this to improve the results and make the module capable of completely replacing LIDAR.

REFERENCES

[1] S. F. Bhat, I. Alhashim, and P. Wonka, “AdaBins: Depth Estimation using Adaptive Bins, ” pp. 1–13, Nov. 2020, [Online]. Available: http://arxiv.org/abs/2011.14141.

[2] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In Computer Vision – ECCV 2012, pages 746– 760, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[3] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. I. J. Robotics Res., 32:1231– 1237, 2013.

[4] L. He, G. Wang, and Z. Hu, “Learning Depth From Single Images With Deep Neural Network Embedding Focal Length,” IEEE Trans. Image Process., vol. 27, no. 9, pp. 4676–4689, Sep. 2018, doi: 10.1109/TIP.2018.2832296.

[5] L. He, G. Wang, and Z. Hu, “Learning Depth From Single Images With Deep Neural Network Embedding Focal Length,” IEEE Trans. Image Process., vol. 27, no. 9, pp. 4676–4689, Sep. 2018, doi: 10.1109/TIP.2018.2832296.

[6] D. Eigen, C. Puhrsch, and R. Fergus, “Depth Map Prediction from a Single Image using a Multi-Scale Deep Network,” Adv. Neural Inf. Process. Syst., vol. 3, no. January, pp. 2366–2374, Jun. 2014, [Online]. Available: http://arxiv.org/abs/1406.2283.