



MONOCULAR DEPTH ESTIMATION USING STATE-OF-THE-ART ALGORITHMS: A REVIEW

Tea Dogandžić*,
Anđela Jovanović

Singidunum University,
Belgrade, Serbia

Abstract:

Monocular Depth Estimation is the process of calculating the depth value of each pixel given a single RGB image. This challenging computer vision task is the main prerequisite for determining scene understanding for applications such as 3D scene reconstruction, augmented and virtual reality. Additionally, a lot of robotics issues, like mapping, localization, and obstacle avoidance for terrestrial and aerial vehicles, depend on depth information. Five monocular depth estimation techniques are compared. This comparison focuses on how generalizable the methods are. According to this study, monocular depth estimation techniques frequently exhibit artifacts when used on images that are not part of the training set, despite performing well on images that are similar to the training images. We test the various approaches using photos that resemble training data as well as paintings or images with odd perspectives.

Keywords:

Depth Estimation, Monocular Depth Estimation, Computer Vision, Deep Learning, Comparison.

Correspondence:

Tea Dogandžić

e-mail:

tea.dogandzic.22@singidunum.ac.rs

INTRODUCTION

Monocular depth estimation (MDE) is a very useful technique in situations where we cannot use sensors and stereo vision to obtain a dense depth map. Instead, we can determine the depth from an ordinary RGB image. It's much cheaper than using expensive equipment and takes less time. In this paper, we compared five methods that are most often used for depth estimation. All models are available online and are free to use and modify. In the first five chapters of this paper, the used algorithms are briefly described and at the end, their applications are presented side by side using different images.



2. ADAPTIVE BINS - ADABINS

Believing that global processing is a lot more powerful when done at high resolution than once the tensors reach a low spatial resolution at or near the bottleneck, the idea of AdaBins is to perform a global statistical analysis of the output of traditional encoder-decoder architecture and to refine the results with a learned post-processing building block that operates at the highest resolution [1]. Authors proposed the analysis and modification of distributions of the depth values. Depth distribution can vary for each image. For example, if we have a picture of an object which is very close to camera, the depth values will be distributed over a very small range. In contrary, if we have a corridor, depth range will go as far as the maximum depth supported by the network. Using AdaBins, depth range is divided into bins where the bin widths changes based on image. The final depth estimation is a linear combination of the bin center values. EfficientNet B5 [2] is used as a backbone for the encoder. With standard feature upsampling decoder [3], it makes a first component of architecture. The second part is proposed adaptive bin-width estimator AdaBins which is based on a transformer [4] that predicts both bin widths and the probability of each pixel to belong to a given bin [1]. Loss function that gave best results is calculated as a summary of Pixel-wise depth loss [5] and Bin-center density loss multiplied by 0.1 for all experiments. Datasets that are used for training are NYU Depth v2 [6], KITTI [7] and SUN RGB-G [8].

3. MIXING DATASETS FOR ZERO-SHOT CROSS-DATASET TRANSFER - MIDAS

The idea behind this approach is to use variety of different datasets for training and testing, because none of the existing datasets is that sufficient and diverse to represent a complete real-world images. In contrary, multiple datasets are complement to each other [9]. Authors made their models and code available on their GitHub (<https://github.com/isl-org/MiDaS>) where you can download pretrained weights. Models were trained on up to 12 datasets (IRS [10], MegaDepth [11], ApolloScape [12], NYU Depth V2 [6], BlendedMVS [13], DIML [14], HRWSI [15], TartanAir [16], ReDWeb [17], 3D Movies, WSVD [18], KITTI [7]) with multi-objective optimization. Experimental protocol used here is zero-shot cross-dataset transfer and it stands for using different datasets for training and testing instead of using subset of one data collection.

There are also used different encoder-decoder architectures such as ResNet-50 [19], ResNeXt-101 [20] or DenseNet-161 [21].

4. DENSE PREDICTION TRANSFORMER

Many of existing architectures for dense prediction are based on CNNs. These architectures are made of encoder, which is based on an image classification network so called “backbone” like ImageNet [22], and decoder that aggregates features from the encoder and converts them to the final dense predictions [23]. This approach has many disadvantages such as the choice of backbone and convolutional image downsampling. There are some work arounds for these issues such as training at higher input resolutions, dilated convolutions [24] or connecting multi-resolution representations in parallel throughout the network [25]. However, convolution is still a bottleneck for prediction quality. Therefore, authors in their work [22] introduced the use of dense prediction transformer (DPT) which changes the base architecture to a transformer based encoder-decoder. Also, they used vision transformer (ViT) [26] as a backbone. There are two models that extend MiDaS with pre-trained weights. The code is available on their GitHub (<https://github.com/isl-org/DPT>).

5. GLOBAL-LOCAL PATH NETWORKS

This method introduces a new architecture and training strategy to further improve the prediction accuracy of the network [27]. GLPDepth uses a transformer based encoder-decoder. The major improvement over other methods is the use of a newly proposed data augmentation trick called “Vertical CutDepth” [28] which enhances the original CutDepth [29] by preserving the vertical geometric information. By maintaining the vertical range of the input image, the network can capture the long-range vertical direction for better prediction [27]. Training loss used in this model is Scale-invariant log scale loss [30]. Datasets NYU Depth v2 [6] and SUN RGB-D [8] are used for validation and evaluation of pre-trained models.

6. ADELAIDDEPTH

AdelaiDepth is an open-source toolbox for monocular depth prediction. Algorithms are available on their GitHub (<https://github.com/aim-uofa/AdelaiDepth>). There are Boosting Depth [31], 3D Scene Shape [32], DiverseDepth [33] [34], Virtual Normal [35], and Depth Estimation Using Deep Convolutional Neural Fields [36] algorithms. In our paper, we will use 3D Scene Shape which offers model weights for ResNet-50 [19] and ResNeXt-101 [20] backbone. The architecture is based on an encoder-decoder, and the network is trained on multiple datasets such as DIML [14], HRWSI [15], Taskonomy [37], 3D Ken Burns [38], and Holopix50K [39]. During training, three losses are combined which are, respectively, Image-level normalized regression loss, Pairwise normal loss, and Multiscale gradient loss.

3D Scene Shape also studies the problem of creating a 3D model from the estimated depth map, and that predicted model is used during the training process in order to regularize the monocular depth prediction module using the surface normals of the estimated 3D shape [28].

7. COMPARISON

In this chapter, we will present our results. We used various pictures in our experiments, but here we will present just a few of them in order to understand the limits and generalization capabilities of each model. We used natural images, synthetic or non-natural, and ambiguous images. The comparison is done by using the magma colormap (see Figure 1). Fine-tuning is not done, because we wanted to simply compare basic methods that are parameters-free.



Figure 1 - Magma Palette.

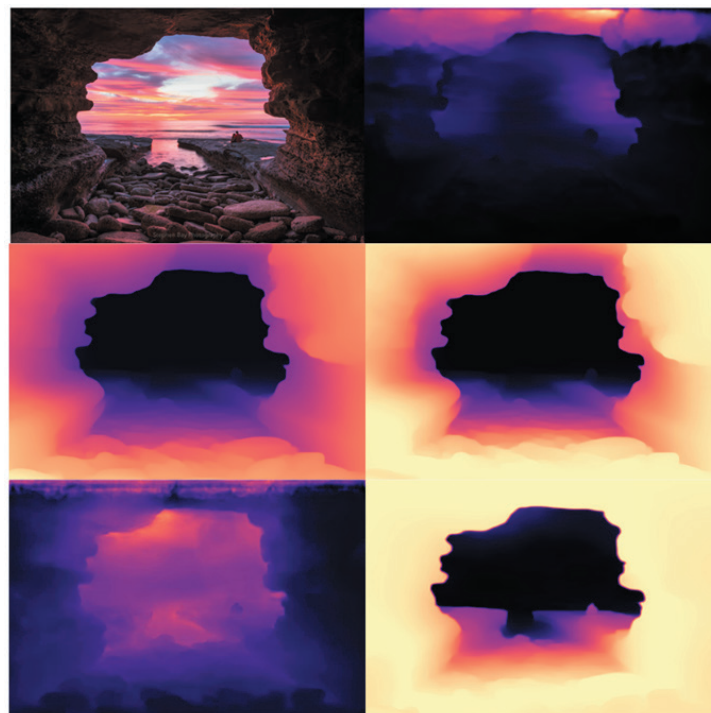


Figure 2 - From top left to bottom right: original image (Cave by Stephen Bay Photography), AdaBins, MiDaS, DPT, GLPDepth, and 3D Scene Shape algorithms.



Figure 3 - From top left to bottom right: original image.

(Trompe l'œil available at <http://fr.pickture.com/blogs/pussycatdreams/1291713-TROMPE-L-OEIL#i>),
AdaBins, MiDaS, DPT, GLPDepth, and 3D Scene Shape algorithms.

8. CONCLUSION

We made a brief overview of existing state-of-the-art algorithms for monocular depth estimation and made a comparison between them using different images. In future work, we would like to investigate other novel approaches and their uses as well as hybrid implementations and combinations of existing ones, fine tuning for better depth predictions, segmentation, and 3D reconstruction from multiple images.

9. REFERENCES

- [1] S. F. Bhat, I. Alhashim and P. Wonka, "AdaBins: Depth Estimation Using Adaptive Bins," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA, 2021.
- [2] M. Tan and Q. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in *International conference on machine learning*, Long Beach, California, USA, 2019.
- [3] I. Alhashim and P. Wonka, "High Quality Monocular Depth Estimation via Transfer Learning," in *CoRR*, 2019.
- [4] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov and S. Zagoruyko, "End-to-End Object Detection with Transformers," in *16th European Conference, Glasgow, UK, 2020*.
- [5] J. H. Lee, M.-K. Han, D. W. Ko and I. H. Suh, "From Big to Small: Multi-Scale Local Planar Guidance for Monocular Depth Estimation," in *arXiv*, 2019.
- [6] N. Silberman, D. Hoiem, P. Kohli and R. Fergus, "Indoor Segmentation and Support Inference from RGBD Images," in *ECCV*, 2012.
- [7] A. Geiger, P. Lenz, C. Stiller and R. Urtasun, "Vision meets Robotics: The KITTI Dataset," *International Journal of Robotics Research (IJRR)*, 2013.
- [8] S. Song, S. Lichtenberg and J. Xiao, "SUN RGB-D: A RGB-D Scene Understanding Benchmark Suite," in *28th IEEE Conference on Computer Vision and Pattern Recognition*, 2015.
- [9] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler and V. Koltun, "Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-Shot Cross-Dataset Transfer," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.
- [10] Q. Wang, S. Zheng, Q. Yan, F. Deng, K. Zhao and X. Chu, "IRS: A Large Naturalistic Indoor Robotics Stereo Dataset to Train Deep Models for Disparity and Surface Normal Estimation," *CoRR*, 2019.
- [11] Z. Li and N. Snavely, "MegaDepth: Learning Single-View Depth Prediction from Internet Photos," *CoRR*, 2018.



- [12] X. Huang, X. Cheng, Q. Geng, B. Cao, D. Zhou, P. Wang, Y. Lin and R. Yang, "The ApolloScape Dataset for Autonomous Driving," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Salt Lake City, UT, USA, 2018.
- [13] Y. Yao, Z. Luo, S. Li, J. Zhang, Y. Ren, L. Zhou, T. Fang and L. Quan, "BlendedMVS: A Large-scale Dataset for Generalized Multi-view Stereo Networks," *CoRR*, 2019.
- [14] Y. Kim, H. Jung, D. Min and K. Sohn, "Deep Monocular Depth Estimation via Integration of Global and Local Predictions," *IEEE Transactions on Image Processing*, vol. 27, pp. 4131-4144, 2018.
- [15] K. Xian, J. Zhang, O. Wang, L. Mai, Z. Lin and Z. Cao, "Structure-Guided Ranking Loss for Single Image Depth Prediction," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 608-617.
- [16] W. Wang, D. Zhu, X. Wang, Y. Hu, Y. Qiu, C. Wang, Y. Hu, A. Kapoor and S. Scherer, "Tartan-Air: A Dataset to Push the Limits of Visual SLAM," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020.
- [17] K. Xian, C. Shen, Z. Cao, H. Lu, Y. Xiao, R. Li and Z. Luo, "Monocular Relative Depth Perception with Web Stereo Data Supervision," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 311-320.
- [18] C. Wang, S. Lucey, F. Perazzi and O. Wang, "Web Stereo Video Supervision for Depth Prediction from Dynamic Scenes," in *2019 International Conference on 3D Vision (3DV)*, 2019, pp. 348-357.
- [19] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," *CoRR*, 2015.
- [20] S. Xie, R. Girshick, P. Dollár and Z. Tu, "Aggregated Residual Transformations for Deep Neural Networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5987-5995.
- [21] G. Huang, Z. Liu and K. Weinberger, "Densely Connected Convolutional Networks," *CoRR*, vol. abs/1608.06993, 2016.
- [22] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248-255.
- [23] R. Ranftl, A. Bochkovskiy and V. Koltun, "Vision Transformers for Dense Prediction," *CoRR*, vol. abs/2103.13413, 2021.
- [24] F. Yu and V. Koltun, "Multi-Scale Context Aggregation by Dilated Convolutions," in *4th International Conference on Learning Representations, ICLR*, San Juan, Puerto Rico, 2016.
- [25] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu and B. Xiao, "Deep High-Resolution Representation Learning for Visual Recognition," *CoRR*, vol. abs/1908.07919, 2019.
- [26] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *CoRR*, vol. abs/2010.11929, 2020.
- [27] D. Kim, W. Ka, P. Ahn, D. Joo, S. Chun and J. Kim, "Global-Local Path Networks for Monocular Depth Estimation with Vertical CutDepth," *CoRR*, vol. abs/2201.07436, 2022.
- [28] T. Ehret, "Monocular Depth Estimation: a Review of the 2022 State," in *Image Processing On Line*, 2023.
- [29] Y. Ishii and T. Yamashita, "CutDepth: Edge-aware Data Augmentation in Depth Estimation," *CoRR*, vol. abs/2107.07684, 2021.
- [30] D. Eigen, C. Puhrsch and R. Fergus, "Depth Map Prediction from a Single Image using a Multi-Scale Deep Network," *CoRR*, vol. abs/1406.2283, 2014.
- [31] G. Xu, W. Yin, H. Chen, C. Shen, K. Cheng, F. Wu and F. Zhao, "Towards 3D Scene Reconstruction from Locally Scale-Aligned Monocular Video Depth," *CoRR*, vol. abs/2202.01470, 2022.
- [32] W. Yin, J. Zhang, O. Wang, S. Niklaus, L. Mai, S. Chen and C. Shen, "Learning to Recover 3D Scene Shape from a Single Image," *CoRR*, vol. abs/2012.09365, 2020.
- [33] W. Yin, Y. Liu and C. Shen, "Virtual Normal: Enforcing Geometric Constraints for Accurate and Robust Depth Prediction," *CoRR*, vol. abs/2103.04216, 2021.
- [34] W. Yin, X. Wang, C. Shen, Y. Liu, Z. Tian, S. Xu, C. Sun and D. Renyin, "DiverseDepth: Affine-invariant Depth Prediction Using Diverse Data," *CoRR*, vol. abs/2002.00569, 2020.
- [35] W. Yin, Y. Liu, C. Shen and Y. Yan, "Enforcing geometric constraints of virtual normal for depth prediction," *CoRR*, vol. abs/1907.12209, 2019.
- [36] F. Liu, C. Shen, G. Lin and I. Reid, "Learning Depth from Single Monocular Images Using Deep Convolutional Neural Fields," *CoRR*, vol. abs/1502.07411, 2015.
- [37] A. Zamir, A. Sax, W. Shen, L. Guibas, J. Malik and S. Savarese, "Taskonomy: Disentangling Task Transfer Learning," *CoRR*, vol. abs/1804.08328, 2018.
- [38] S. Niklaus, L. Mai, J. Yang and F. Liu, "3D Ken Burns Effect from a Single Image," *CoRR*, vol. abs/1909.05483, 2019.
- [39] K. Xian, J. Zhang, O. Wang, L. Mai, Z. Lin and Z. Cao, "Structure-Guided Ranking Loss for Single Image Depth Prediction," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2020, pp. 608-617.