Towards Explainability in Monocular Depth Estimation

Vasileios Arampatzakis[1,2][0000−0003−4320−3740], George Pavlidis[2][0000−0002−9909−1584], Kyriakos Pantoglou[2][0009−0008−5683−3382], Nikolaos Mitianoudis[1][0000−0003−0898−6102], and Nikos Papamarkos[1][0000−0003−2730−0006]

[1] Democritus University of Thrace, Xanthi, Greece
{vaarampa, nmitiano, papamark}@ee.duth.gr
[2] Athena Research Center, Xanthi, Greece
{vasilis.arampatzakis, kyriakos.pantoglou, gpavlid}@uathenarc.gr

Abstract. The estimation of depth in two-dimensional images has long been a challenging and extensively studied subject in computer vision. Recently, significant progress has been made with the emergence of Deep Learning-based approaches, which have proven highly successful. This paper focuses on the explainability in monocular depth estimation methods, in terms of how humans perceive depth. This preliminary study emphasizes on one of the most significant visual cues, the relative size, which is prominent in almost all viewed images. We designed a specific experiment to mimic the experiments in humans and have tested state-of-the-art methods to indirectly assess the explainability in the context defined. In addition, we observed that measuring the accuracy required further attention and a particular approach is proposed to this end. The results show that a mean accuracy of around 77% across methods is achieved, with some of the methods performing markedly better, thus, indirectly revealing their corresponding potential to uncover monocular depth cues, like relative size.

Keywords: computer vision · monocular depth estimation · explainability.

1 Introduction

Research by Nagata [1], further classified by Cutting and Vishton [2], presented a complete set of visual depth cues, as a result of specifically designed experiments in humans. These cues include *occlusion*, *relative size*, *relative density*, *height in the visual field*, *aerial perspective*, *motion perspective*, *convergence*, *accommodation*, and *binocular disparity*. The combination of these visual depth cues appears to be associated with the comprehensive understanding of a scene. Later, other researchers tried to systematically review the domain and present a more thorough view of depth estimation in humans [3].

The scientific community has long faced a significant challenge in achieving depth perception in mechanical systems. The accurate estimation of depth is a

crucial task in machine visual perception. Depth estimation involves reconstructing the missing dimension, which represents the distance between the objects and the observer in a three-dimensional (3D) scene, through a two-dimensional (2D) projection of the scene.

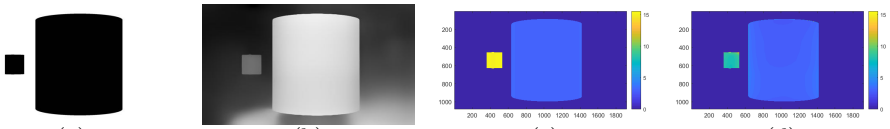
Nevertheless, there is no study, to the best of our knowledge, that clearly connects depth cues, as defined for humans, with the ability of modern depth estimation methods to estimate depth in monocular 2D images, thus affecting the explainability of those methods. In this paper, we explore towards addressing this issue, by using specifically designed data, and focus on the relative size depth cue, a prominent cue in projected scenes. We evaluate selected state-of-the-art depth estimation methods using those data and provide insights into their explainability, within our context.

2 Deep Learning-Based Monocular Depth Estimation

Traditional methods relied on assumptions, constraints, and optimizations to provide detailed depth estimates. However, these methods faced limitations such as a restricted measurement range, sensitivity to outdoor lighting conditions, calibration requirements, and high energy consumption, which hindered the utilization of sensor-based techniques involving RGB-D and LiDAR sensors. Additionally, approaches based on image pairs or sequences could only calculate depth values for sparse points.

To tackle these challenges, researchers proposed the usage of deep learning. Deep Learning methods achieved high performance in estimating dense depth maps. In tasks like depth calculation with high complexity, where it is nearly impossible to apply classical pattern recognition approaches, deep learning methods achieved remarkable results. In the following study we focus on the significant work by Ibraheem et al. [4], Godard et al. [5], and Ranftl et al. [6], who made important contributions to the field. Ibraheem et al. [4] proposed a novel approach (**DenseDepth**) to estimate high-resolution depth maps from RGB images using a transfer learning-based encoder-decoder network. The encoder was a pretrained DenseNet-169, fine-tuned on NYU Depth v2[3] and KITTI datasets[4]. The authors reported state-of-the-art results in typical and qualitative aspects and in terms of generalisation. Godard et al. [5] focused on estimating depth using video sequences, stereo pairs, or a combination of both (**Monodepth2**). They introduced several improvements. The system was trained on subsets of the KITTI dataset. The authors reported state-of-the-art results, outperforming other methods at that period. Ranftl et al. [6] proposed a novel approach (**MiDaS**) to enhance the robustness of depth estimation models and address the challenge of dataset bias. The authors reported results outperforming previous methods and particularly in terms of generalisation, making their method one of the most effective to date.

[3] NYU-v2, indoor images (rooms & hallways scenes) [7].
[4] KITTI, outdoor images (urban & street scenes) [8].



(a) (b) (c) (d)

Fig. 1. Process example: (a) original image, (b) predicted depth, (c) pseudo-coloured groundtruth depth, (d) pseudo-coloured masked predicted depth.

3 The explainability experiment and results

To create meaningful explainability experiments we tried to mimic the relevant work done in humans. In the original experiments, as described in the Introduction, humans were asked to assess the relative distance of partially viewed untextured objects against neutral backgrounds. Thus, a particular artificial dataset had to be created, using 3D modeling software. The new dataset consists of 23800 2D images of black cylindrical objects at various distances against a white background, created through perspective projections of the corresponding virtual 3D scenes. Those data were used to test the three selected pretrained state-of-the-art methods. The core idea is to indirectly assess the explainability of the methods in learning the relative size cue, by providing test examples which contain only this single cue. We considered all published variations of the considered models. Furthermore, to evaluate the models, binary masks were applied to focus only on the pixels associated with objects in the scene. In addition, we adopted the scale and shift pre-alignment suggested by [6]. The assessment of depth predictions was based on common error and accuracy metrics, the most popular of which were introduced by Eigen [9]; we used those definitions. An indicative example of the process is shown in Fig. 1, where an original image of two objects at different distance is shown in (a), the predicted depth in (b), and pseudo-coloured representations of the groundtruth (c) and the predicted depth (d).

Additionally, we observed that the size of the objects in the images plays a significant role in the estimates of the error and accuracy. In the example shown, an error in the estimated depth of the far object (depicted significantly smaller) will have a negligible impact on the metrics, while the estimate should be balanced. This observation led to the introduction. After the objects become equally sized, the metrics are calculated, ensuring equal significance in error estimates for both objects in the scene. Fig. 2 depicts the overall average results. The gray bars represent results obtained by using the metrics in the typical way, whereas the black bars represent results obtained after the rescaling process. As expected, rescaling the smaller object results in lowering the accuracy (increasing the error) on the average.

Apparently, the three first variations of MiDaS outperform the other methods, particularly when using rescaling, and achieve an average of $\delta_1 = 0.85$ (85%). The same is reflected in the error estimates. From the experiments, it seems possible that MiDaS partially learns the relative size cue, although more experiments are needed to generalise this remark. In addition, the DenseDepth

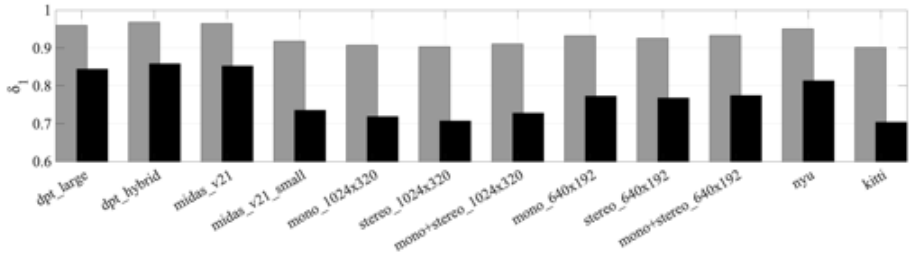


Fig. 2. The accuracy with a threshold δ_1 of the pertained models tested on our dataset.

method pretrained on NYU-v2 also exhibits increased accuracy compared with MiDaS, whereas the same model pretrained on KITTI is among the worst cases. This shows that the training dataset plays a significant role in learning depth cues and this should be considered in situations where a connection with explainable results is required.

This is a preliminary study, focusing on a single monocular depth cue, the relative size, and only a small set of methods. Explainability in depth estimation was considered on the basis of how humans estimate depth and a connection between the visual depth cues was attempted with the depth estimation provided by state-of-the-art approaches. To enable this, a new dataset was created, to mimic the original experiments in humans. This meant that the relative size cue should be isolated and no other depth cues should be present in the images. This study is ongoing and more datasets are created for each of the visual depth cues to assess the effectiveness of the existing methods, and thus, to conclude on the potential explainability in learning those cues. Overall, the final dataset will become a benchmark for testing the explainability of depth estimation methods.

4 Conclusion

In this study, we tried to approach explainability in depth estimation deep learning methods in terms of human perception. To this end, a specific visual depth cue was selected (the relative size) and a new dataset was created to mimic the experiments in humans. Three state-of-the-art pretrained methods were selected and tested against this dataset. As this dataset is limited to provide only a single cue, the accuracy of the methods indirectly reflects their success in learning the selected depth cue. In addition, it has been observed that the typical assessment metrics should be applied on rescaled versions of the image objects, in order to balance the estimated accuracy. Overall, the methods returned interesting accuracy results. Currently, we are expanding the dataset to include other visual depth cues and design new experiments to evaluate the efficiency of state-of-the-art methods.

References

1. S Nagata. How to reinforce perception of depth in single two-dimensional pictures. *Spatial displays and spatial instruments*, 29, 1987.
2. James E Cutting and Peter M Vishton. Perceiving layout and knowing distances: The integration, relative potency, and contextual use of different information about depth. In *Perception of space and motion*, pages 69–117. Elsevier, 1995.
3. Ian P Howard. *Perceiving in depth, volume 1: basic mechanisms*. Oxford University Press, 2012.
4. Ibraheem Alhashim and Peter Wonka. High quality monocular depth estimation via transfer learning. *arXiv preprint arXiv:1812.11941*, 2018.
5. Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 3828–3838, 2019.
6. René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 2020.
7. Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European conference on computer vision*, pages 746–760. Springer, 2012.
8. Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361. IEEE, 2012.
9. David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in neural information processing systems*, pages 2366–2374, 2014.