Pacific Graphics (2021)
M. Okabe, S. Lee, B. Wuensche and S. Zollmann (Editors)

User-centred Depth Estimation Benchmarking for VR Content Creation from Single Images

Anthony Dickson [1] Alistair Knott[1] Stefanie Zollmann[1]

[1]University of Otago, New Zealand

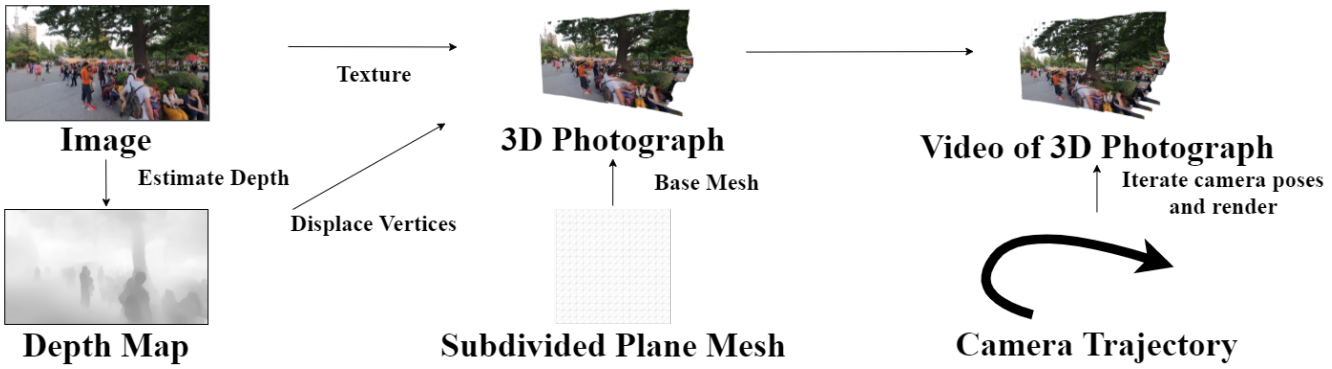

Figure 1: *The process for creating 3D photographs from 2D images. The 3D photographs are created by taking a RGB-D image and texturing and displacing the vertices on a subdivided plane mesh. A 3D photograph dataset is then created by using a predefined camera trajectory.*

Abstract

The capture and creation of 3D content from a device equipped with just a single RGB camera has a wide range of applications ranging from 3D photographs and panoramas to 3D video. Many of these methods rely on depth estimation models to provide the necessary 3D data, mainly neural network models. However, the metrics used to evaluate these models can be difficult to interpret and to relate to the quality of 3D/VR content derived from these models. In this work, we explore the relationship between the widely used depth estimation metrics, image similarly metrics applied to synthesised novel viewpoints, and user perception of quality and similarity on these novel viewpoints. Our results indicate that the standard metrics are indeed a good indicator of 3D quality, and that they correlate with human judgements and other metrics that are designed to follow human judgements.

CCS Concepts

• ***General and reference*** → ***Evaluation;*** • ***Computing methodologies*** → *Computer graphics; Neural networks; Computer vision;*

1. Introduction

Estimating depth from a single image without additional sensors is a challenging task that has applications in many areas ranging from driver-less cars to image editing. Recently, depth estimation from single images has been used for 3D content creation such as for creating 3D photographs [KMA*20].

Depth estimation models are commonly evaluated with the following metrics:

- accuracy under a threshold: the percentage of the estimated depth $\hat{d}_i$ such that $\max(\frac{\hat{d}_i}{d_i}, \frac{d_i}{\hat{d}_i}) = \delta <$ threshold for threshold values of 1.25, 1.25^2 and 1.25^3 (we refer these to as DEL1, DEL2 and DEL3 in the results)
- mean relative error (REL) $\frac{1}{N}\sum_{i=1}^{N}\frac{||d_i-\hat{d}_i||_1}{d_i}$
- root mean squared error (RMSE) $\sqrt{\frac{1}{N}\sum_{i=1}^{N}(d_i-\hat{d}_i)^2}$.

The main reason for using these metrics is that that an early focus of monocular depth estimation (MDE) was on autonomous navi-

gation. However, with more interest on using these models for 3D content creation, an important question arises: Are these traditional benchmarks suitable for evaluating depth estimation models for 3D and VR content creation?

2. Method

We evaluate a set of depth estimation models on a dataset of 3D photographs created from the NYU depth dataset [SHKF12]. The 3D photographs are created by taking a RGB-D image and texturing and displacing the vertices on a subdivided plane mesh as illustrated in Figure 1. We use this approach with five sources of depth data:

1. Ground truth depth data that is included with the NYU dataset.
2. Large depth estimation model (about 160 million parameters) [HOZO19].
3. A small model (6 million parameters) based on (2) where the SENet encoder has been swapped out for EfficientNet-B0 [TL19].
4. A untrained version of the small model (3) that serves as a baseline for low quality.
5. Flat depth maps (i.e. no depth) which are used to verify whether users can detect the presence of 3D effect.

We quantitatively compare novel viewpoints from the 3D photographs using the following image similarity metrics: PSNR, SSIM [WSB03], LPIPS [ZIE*18] and reprojection error. Here the reprojection error is the mean Euclidean distance between matching SIFT features in a pair of images.

In addition, we designed a user study to evaluate the quality from a user's perspective. In our user study, we recruited 20 participants through Amazon's Mechanical Turk service. Participants were presented with videos of 4 scenes from the dataset in pairs (one of the ground truth, one from the other models) and asked for ratings of how similar the two videos are and the quality of the 3D effect in each video.

3. Results

Overall, our results suggest that the standard depth estimation metrics are indeed a good indicator of 3D quality. The results from the depth estimation metrics are within expectations with the untrained model resulting in a relative error of 0.991, the small model 0.141 and the large model 0.115. These results are consistent with the quantitative results on the 3D photograph dataset, where the untrained model scored 0.675 for SSIM, the small model 0.696 and the large model 0.757. The depth estimation results also are consistent with the user study results. When we asked participants how much they agree/disagree with the statement "The 3D effect in this video is realistic." on a 7-point likert scale (1 - strongly disagree, 4 - neutral, 7 strongly agree), the median rating was 2 for the untrained model, 4 for the small model and 5 for the large model. Statistical analysis indicated a significant effect of MDE model on realism.

4. Conclusion and Future Work

We presented our results on using a user-centred approach for benchmarking monocular depth estimation methods for VR content creation from single images. In order to do so we created a dataset that renders 3D meshes of scenes from novel viewpoints. We analysed these images with image-based metrics and presented these images to users in a user study to analyse if they see differences in quality. Our experiments showed that these metrics can be used for quality assessments of depth estimation methods. However we also found that some metrics, such as SSIM and PSNR are not reliable enough to find difference that would be obvious to human observers (Figure 2). For future work, it would be important to evaluate a larger range of depth estimation models on a wider variety of scenes. The mesh creation process produces content with obvious artefacts and adopting existing approaches to 3D photograph creation would help improve the quality of the dataset and the user study results. We hope that our work contributes to improving the quality assessments and benchmarking of depth estimation methods in particular for their suitability of content creation for VR.

Figure 2: *A 3D photograph rendered from a novel viewpoint using depth maps from: the small model with random weights (left) and the trained small model (right). In the bottom right corner of each image are the PSNR and SSIM metrics for those images.*

References

[HOZO19] HU J., OZAY M., ZHANG Y., OKATANI T.: Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries. In *Proc. - 2019 IEEE Winter Conf. Appl. Comput. Vision, WACV 2019* (2019), pp. 1043–1051. arXiv:1803.08673, doi:10.1109/WACV.2019.00116. 2

[KMA*20] KOPF J., MATZEN K., ALSISAN S., QUIGLEY O., GE F., CHONG Y., PATTERSON J., FRAHM J.-M., WU S., YU M., ZHANG P., HE Z., VAJDA P., SARAF A., COHEN M.: One shot 3d photography. *ACM Trans. Graph. 39*, 4 (July 2020). doi:10.1145/3386569.3392420. 1

[SHKF12] SILBERMAN N., HOIEM D., KOHLI P., FERGUS R.: Indoor segmentation and support inference from RGBD images. In *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)* (2012), vol. 7576 LNCS, pp. 746–760. doi:10.1007/978-3-642-33715-4_54. 2

[TL19] TAN M., LE Q. V.: Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv preprint arXiv:1905.11946* (2019). 2

[WSB03] WANG Z., SIMONCELLI E. P., BOVIK A. C.: Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003* (2003), vol. 2, Ieee, pp. 1398–1402. 2

[ZIE*18] ZHANG R., ISOLA P., EFROS A. A., SHECHTMAN E., WANG O.: The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2018), pp. 586–595. 2