Benchmarking Monocular Depth Estimation Models for VR Content Creation from a User Perspective

Anthony Dickson
Department of Computer Science
University of Otago
Dunedin, New Zealand
dican732@student.otago.ac.nz

Alistair Knott
Department of Computer Science
University of Otago
Dunedin, New Zealand
alik@cs.otago.ac.nz

Stefanie Zollmann
Department of Computer Science
University of Otago
Dunedin, New Zealand
stefanie.zollmann@otago.ac.nz

***Abstract*—Exploring monocular images and videos in 3D in Virtual Reality (VR) requires reliable methods for estimating depth as this has not been captured in the original footage. There are monocular depth estimation methods that address this requirement by using neural network models to predict depth values for each pixel. However, as this is a quickly moving field and the state-of-the-art is constantly evolving, how do we choose the best depth estimation model for 3D content creation for VR? It can be difficult to interpret the widely used benchmark metrics and how they relate to the quality of the created 3D/VR content.**

In this paper, we explore a user-centred approach to evaluating depth estimation models with regard to 3D content creation. We look at evaluating these models based on content created with these models that an end-user might see, rather than just their immediate output (depth maps) that does not directly contribute to users' perceptual experience. In particular, we investigate the relationship between commonly used depth estimation metrics, image similarity metrics applied to synthesised novel viewpoints, and user perception of quality and similarity of these novel viewpoints. Our results suggest that the standard depth estimation metrics are indeed good indicators of 3D quality, and that they correspond well with human judgements and image similarity metrics on novel viewpoints synthesised from a range of sources of depth data. We also show that users rate a state-of-the-art depth estimation model as almost visually indistinguishable from the outputs derived from ground truth sensor data.

I. Introduction

Monocular depth estimation methods provide the ability to create 3D content from single images. This creates several new exciting opportunities such as creating 3D photographs [1] from a single monocular photograph, 3D Panoramas [2] as well as VR content [3]. In recent years, many novel approaches for computing depth from single images have been proposed and have advanced the state of the art. In particular, monocular depth estimation (MDE) models that employ neural networks have demonstrated their ability to accurately reconstruct depth maps from single images. There have been approaches that have engineered the architecture to allow the training of larger and deeper networks [4], ones that have cast the problem of depth estimation as an ordinal regression problem [5], ones that have included geometric priors [6], ones that have been trained on a wide range of datasets [7] and even ones that forego the standard fully convolutional architecture in favour of a visual transformer architecture [8].

It is standard practice to benchmark newly proposed models against the state-of-the-art by using ground truth depth data [9], [10] as well as using a standard set of metrics that often include at least: relative error, RMSE, and thresholded accuracy at three thresholds (1.25, 1.25^2 and 1.25^3). One thing in common with all of these metrics is that they measure depth error on a pixel-wise basis. The main reason for using these metrics is that an early focus of MDE was on autonomous navigation. However, with more interest in using MDE for 3D content creation, an important question arises: Are these traditional benchmarks suitable for benchmarking MDE models for 3D and VR content creation? In fact, these metrics can be rather difficult to interpret in this sense. Imagine a 3D application that uses MDE models in the 3D content creation process and we present a user with the outputs from two different models, what would the differences in the metrics mean for the user? How do the metrics relate to the quality of what the user sees? And how do we interpret the absolute figures from these metrics (e.g. what does a relative error of 0.150 mean)? Is there a point where the metrics relate to an acceptable quality for 3D content?

In this paper, we work towards answering these questions and present the full results of our poster on this topic [11]. We first generate a dataset of animated novel viewpoint videos from the NYU dataset [9] by using different MDE models. We then compare the frames of these videos with frames rendered with ground truth depth using widely used image similarity metrics. We then conduct a user study where participants are tasked with comparing pairs of videos in terms of similarity and rating individual videos in terms of quality. Our results indicate that the standard depth estimation metrics behave similarly to human judgements and to image-centric metrics. We also show that traditional image similarity metrics can struggle with evaluating 3D deformations such as in Figure 1.

II. Background

To the best of our knowledge, there is little work that examines the ability of the standard depth estimation metrics to accurately and thoroughly evaluate depth estimation models for VR content creation.

978-1-6654-0645-1/21/$31.00 ©2021 IEEE

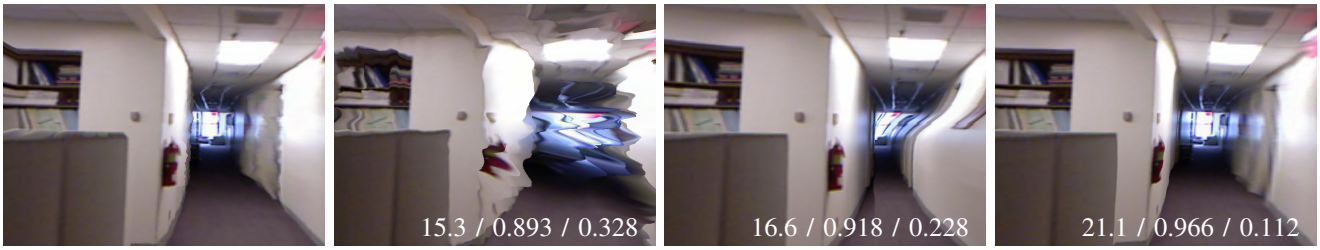

Fig. 1. An image from the NYU dataset rendered from different perspectives using depth maps from: the ground truth data, a small depth estimation model with random weights (HU-ENB0-RW), a small depth estimation model (HU-ENB0) and a large depth estimation model (HU-SENET). In the bottom right corner of each synthesised image are the PSNR, SSIM and LPIPS metrics for each image compared to the ground truth. Notice how PSNR and SSIM give similar numbers for the model with random weights HU-ENB0-RW (middle-left) and the same model that has been trained HU-ENB0 (middle-right), despite the obvious and large difference in quality.

A. Standard Benchmark Metrics

For clarity, we will define the metrics that we refer to as 'standard benchmark metrics'. These are the metrics that are common throughout the vast majority of the recent literature. These metrics are:

- Accuracy under a threshold: percentage of $\hat{d}_i$ such that $\max(\frac{\hat{d}_i}{d_i}, \frac{d_i}{\hat{d}_i}) = \delta <$ threshold where: d_i is the depth value at pixel i of the ground truth depth map d; and $\hat{d}_i$ is the depth value at pixel i of the estimated depth map $\hat{d}$. In depth estimation literature, it is common practice to use the following three threshold values: $\delta < 1.25$, $\delta < 1.25^2$ and $\delta < 1.25^3$. We refer to these as DEL1, DEL2 and DEL3 in the results.
- Mean relative error (REL): $\frac{1}{N}\sum_{i=1}^{N}\frac{||d_i - \hat{d}_i||_1}{d_i}$ where N denotes the total number of valid pixels (pixels with nonzero depth) for a given sample.
- Root mean squared error (RMSE): $\sqrt{\frac{1}{N}\sum_{i=1}^{N}(d_i - \hat{d}_i)^2}$

It is unclear whether these metrics are enough to evaluate these models, especially if the application of the models goes beyond just depth estimation. What if we want to use these models for 3D content creation? Is choosing the model with the best performance on the commonly used metrics enough for this? One challenge with these metrics is that they are calculated on pixel-wise residuals without explicit regard to the spatial relationships between pixels.

B. Evaluating Depth Estimation Metrics

Cadena et al. [12] discuss the limitations of the standard metrics due to them being 'image-space' metrics. They bring up three main issues: resolution, density and coverage. Many methods, including state-of-the-art methods, predict depth maps at a lower resolution than that of the ground truth and thus requiring either the estimated or ground truth to be rescaled. Usually, the estimated depth map is up-scaled via bilinear interpolation to match the resolution of the ground truth, and the authors argue that the interpolated data often disagrees with the ground truth data, negatively impacting the evaluation. Density and coverage are closely related; the issue with density is that methods that predict sparse depth may be at an advantage over methods that predict dense depth in current evaluation methods, whereas coverage is more concerned with methods that predict dense depth for only a portion of the scene (e.g. planar surfaces such as roads). The authors propose that rather than evaluating depth estimation models in image space, we can evaluate them in 3D space by projecting points into 3D space using the depth map and known camera parameters.

Koch et al. [13] posit that MDE models are evaluated with metrics that compare global statistics of depth residuals and that these metrics do not directly assess the accuracy of planar surfaces or depth discontinuities (boundaries). The accuracy of planar surfaces is indeed an issue with MDE models as demonstrated in Figure 1 and is an issue that Niklaus et al. [1] had to directly address. Koch et al. propose four metrics: two that measure the smoothness and orientation of estimated 3D planes and ground truth 3D planes (they create a dataset with annotations of planar surfaces); and two that measure the accuracy and completeness of depth boundaries.

These papers only look at how depth estimation models are evaluated for the task of depth estimation and do not directly look into how the evaluation methods relate to the quality of the 3D content created with these models. Cadena et al.'s approach uses 3D point clouds to evaluate depth estimation models [12], but they do not look into synthesised novel views whose quality is very dependent on the quality of the depth data. Koch et al. [13] improve upon the standard metrics, but they do not look into evaluating depth estimation models in 3D space or 3D content created with these models. In this paper, we create a 3D mesh and evaluate images rendered from this mesh with image similarity metrics and human judgements.

III. Dataset Creation

To evaluate depth estimation models for 3D content creation, we create a video dataset that simulates different perspectives from a single image. We create a video from each of the images in the test set of the NYU v2 dataset [9]. Each video is created from a single RGB-D frame and shows a virtual camera being moved around a 3D mesh created from this RGB-D frame.

A. From RGB-D Frame to Textured Triangle Mesh

As input, we use a RGB-D frame where the depth map can be either be the ground truth depth data or estimated by an MDE model. Depth maps are normalised to the range [0, 1] to simplify the handling of depth maps of different scales and ranges. The RGB-D frame is used to create a textured triangle mesh. We start with a plane mesh and subdivide it some number times depending on how dense we want the mesh to be. In our experiments, we create a plane mesh on the XY plane (y-up) with 257 vertices along the sides of the mesh (total of $\sim$130K triangles). The y-position of the vertices are adjusted so that the plane mesh has the same aspect ratio as the input. Since the vertex positions are in normalised device coordinates, they range from 0 to 1 and adjusting the mesh to the same aspect ratio can be done by multiplying the y-coordinates by the height of the input images divided by the width of the input images. The z-coordinates of the vertices are displaced by the depth value of the corresponding pixel in the depth map, and the colour of the vertices are set in a similar way by sampling the input colour frame (using nearest-neighbour sampling).

B. Rendering

We render the meshes created from RGB-D frames with OpenGL. In addition to the meshes, we also need to define a virtual camera and a camera trajectory. The virtual camera uses a projection matrix created with a vertical field of view of 18 degrees and an aspect ratio based on the input images (4:3 for the NYU dataset). We choose a small field of view so that we can move the camera around while showing the edges of the mesh where there is no data as little as possible. For the camera trajectory, we move around the mesh in an elliptical path and the camera is rotated towards the centre of the mesh in order to help emphasise the 3D effect (Figure 2). Since we are using a virtual camera with different parameters to the camera(s) used to capture the RGB-D data and normalised depth values, the meshes may exhibit a 3D effect that is either too pronounced or too subtle. As such, the z-coordinates of the mesh vertices are also multiplied by a scalar to reach the desired level of 3D effect. For the NYU dataset, we found that a value of 4.0 gave the most accurate results given the virtual camera parameters. We record the sequence at 60 FPS and save the output to a video file. The rendering process described above is repeated for each of the models listed below.

C. Depth Estimation Models

For our experiments, we use five different models/depth map sources:

- GT: The ground truth depth maps from the NYU dataset.
- FLAT: This model simply returns a depth map where all pixels are set to zero and is only used in the user study. The reason for including this model is explained in Section V.
- HU-SENET: The state-of-the-art MDE model from Hu et al. [14]. We include this model as a representative of the state-of-the-art.

Fig. 2. Four synthesised frames using the ground truth depth maps showing the camera trajectory (left to right, top to bottom).

- HU-ENB0: A version of the model architecture from Hu et al.'s paper where we use EfficientNet-B0 for the encoder portion of the network. This model has 5.3 million parameters and is much smaller than HU-SENET which has 157 million parameters. It also uses around half the memory but it is less accurate.
- HU-ENB0-RW: An untrained version of the HU-ENB0 that uses randomly initialised weights serving as a low quality baseline.

D. Limitations

The way that meshes are created is simple, and it may result in artefacts. Mainly, we see stretching along strong edges and blurring of the textures along these edges. The solution, which is implemented in previous works [1], [2], is to either stick with a point cloud representation, to tear the mesh along strong edges or to use a layered representation.

IV. Image Similarity Benchmark

We employ widely used image similarity metrics to generate quantitative assessments over the entire dataset. The purpose of this benchmark is to provide a way to evaluate the generated dataset in a way that follows human judgements using quantitative measures. It also allows us to evaluate the entire dataset, which would be difficult, not to mention costly, if it were to be done with human judgements. It also provides some insight into the strengths and weaknesses of the widely used image similarity metrics.

We use four metrics for quantitatively measuring image similarity: SSIM, PSNR, LPIPS and MIFD. SSIM [15] and PSNR [16] are ubiquitous in image similarity tasks that compare statistics on the pixel values of two images. LPIPS [17] is a newer metric that has seen quick adoption in recent work. It takes a different approach to comparing images by comparing the outputs from the various layers of a convolutional neural network. Mean Image Feature Distance (MIFD) is similar to reprojection errors used in camera calibration and measures how far image features have moved from one image to another. More precisely, we calculate MIFD as the average Euclidean

distance between SURF features that are present in both images with the SURF features being filtered using Lowe's ratio test with a threshold of 0.7. For the details of the other metrics, we direct readers to the cited sources.

These metrics all have their own pros and cons. For example, PSNR is a simple metric that compares the observed difference (as MSE) and the maximum potential difference between two images. SSIM is a bit more sophisticated in how it compares two images and compares various statistics on a sliding window of pixels from the images. LPIPS is more complicated to compute than PSNR or SSIM, but has been shown to more closely follow human judgements when comparing an image to a copy that has basic transformations applied to it (e.g. blur). MIFD is different from the other metrics in that it is more focused on geometric differences as opposed to differences in pixel values.

For this benchmark, we evaluate five evenly spaced frames per video starting with the first frame as shown in Figure 2. This cuts down the total number of frames from 192K to 3.2K which makes the benchmark close to 60 times faster to run. This is especially important since MIFD relies on image feature detection and matching algorithms that cannot be applied to batches of images like the other metrics and is thus much slower to compute on large datasets.

V. User Study

In order to compare how users perceive 3D content and how their perception relates to the standard benchmark metrics, we run a user study to evaluate user perception of the created dataset. We asked participants to assess single videos in terms of absolute quality and pairs of videos in terms of similarity. We recruited 20 participants through Amazon's online service Mechanical Turk (MTurk). We allotted each participant up to 20 minutes to ensure the results were not affected by the fatigue/boredom of the participants. This time limit also limits how many different conditions (models and scenes) we can include in the study. The NYU test set consists of 654 RGB-D frames and it would be impractical and costly to run a study using all of these data points. Hence, we choose to use just four of the images (see Figure 3) that have a wide range of depth values and relatively well lit. The study was approved by the University of Otago Human Ethics Committee.

Each of the models used in the dataset generation are paired off with the ground truth and each their videos are combined into a new video that shows them side-by-by. With four scenes and four model pairs, we have a total of 16 different videos for each participant to evaluate. The order of videos was randomised using Latin square. We refer to each paired video as a 'task', and we ask three questions per task. First, we ask participants to rate how similar the inputs are, then we ask them to rate how realistic the 3D effect is in each of the inputs (two) using a 7-point Likert score.

Two of the models, FLAT and HU-ENB0-RW, are included to ensure that participants are paying attention and understand the tasks they have been given. The flat depth maps are there to check whether participants are able to distinguish between

Fig. 3. The scenes used in the user study.

videos with and without a visible 3D effect. Participants are coached to give a lower rating for realism if they notice flat-looking geometry/videos. The depth maps based on HU-ENB0-RW and the resulting rendered novel viewpoints are low quality and easy to distinguish as unrealistic. If a participant does not consistently rate the outputs of this model as lower quality and dissimilar to the ground truth, it would indicate that they likely do not understand the task or are not paying attention and just choosing arbitrary answers. To ensure that participants spend enough time to carefully answer the questions, we program the study's webpage such that once the participant starts a video, it must be played at least three times before participants can fill out their responses.

Participants are given some brief training/coaching before they start. They are shown examples of what could be considered realistic versus what could be considered unrealistic (e.g. curved walls, noisy depth). They are also shown examples of a scene rendered with a flat depth map versus one rendered with the ground truth to exemplify what a lack of depth may look like. While giving participants coaching on what to look out for could potentially introduce unwanted bias, participants seemed to struggle to evaluate realism in early tests where they were given minimal instruction. We also collected demographic data from the participants (12 male, 4 female, age ranging from 20-49, median=32, VR usage ranging from never (5%), once or twice (30%), at least once a year (25%), at least once a month (15%), to at least once a week (25). All participants reported to have normal/corrected-to-normal vision.

VI. Results

In the following, we will present the results from the depth estimation benchmark, the image similarity benchmark and the user study and discuss them.

A. Depth Estimation Benchmark

Overall the results from the standard depth estimation benchmark are what one would expect; the larger model (HU-SENET) is more accurate than the smaller model (HU-ENB0) and the model with random weights (HU-ENB0-RW) is very

TABLE I
BENCHMARK RESULTS USING THE STANDARD METRICS.

	↓ Lower is Better		↑ Higher is Better		
Model	REL	RMSE	DEL1	DEL2	DEL3
HU-ENB0-RW	0.991	3.063	0.000	0.000	0.000
HU-ENB0	0.141	0.608	0.812	0.959	0.990
HU-SENET	**0.115**	**0.530**	**0.866**	**0.975**	**0.993**

TABLE II
IMAGE SIMILARITY BENCHMARK RESULTS ON ENTIRE NYU TEST SET.

	SSIM	PSNR	LPIPS	MIFD
HU-ENB0-RW	0.675	16.1	0.453	93.9
HU-ENB0	0.696	16.7	0.346	40.4
HU-SENET	**0.757**	**19.1**	**0.249**	**26.4**

inaccurate (Table I). However, it is important to note that some of the results between HU-SENET and HU-ENB0 only differ by a small amount (e.g. REL 0.115 vs 0.141).

B. Image Similarity Benchmark

Based on the image metrics, HU-SENET also outperforms the other models (Table II). One area where the results differ from the depth estimation benchmark results is how well the metrics discriminate between the models HU-ENB0 and HU-ENB0-RW. In the depth estimation metrics, we observe a difference of between 5-6x for REL and RMSE on these models, whereas SSIM only differs by 3%, PSNR by 3.6%, LPIPS by 24%. This is quite surprising considering how distorted and noisy the outputs from HU-ENB0-RW are compared to HU-ENB0 (refer back to Figure 1 for an example). The metric that shows the largest difference between these models is MIFD which is 2.3x more for HU-ENB0-RW than HU-ENB0. Overall, LPIPS and MIFD seem to have more discriminative power than SSIM or PSNR. MIFD seems to be the most robust measure here and this is likely due to it being based around geometric error, however it is slower to compute since it is difficult to compute on images batches like the other metrics.

C. User Study

The results from the user study resemble the ranking of the two previous benchmarks where the largest model ranks the highest followed by HU-ENB0 and then HU-ENB0-RW for both measurements similarity and realism. However, it is important to note HU-ENB0 and FLAT score similar neutral results for both measurements.

A Kruskal-Wallis rank sum test indicated a main effect of MDE model on similarity rating (Figure 4). Pairwise comparisons using Wilcoxon rank sum test with Bonferroni correction indicated a significant difference between the randomised weights GT-HU-ENB0-RW (MD=4) and all other models: GT-FLAT (MD=5, p=0.0001), GT-HU-ENB0 (MD=5, p=7.67e-06) and GT-HU-SENET (MD=6 , p=4.8e-10) (Figure 5). We also found a significant difference between GT-FLAT and both GT-HU-ENB0-RW (p=0.0001) and GT-HU-SENET

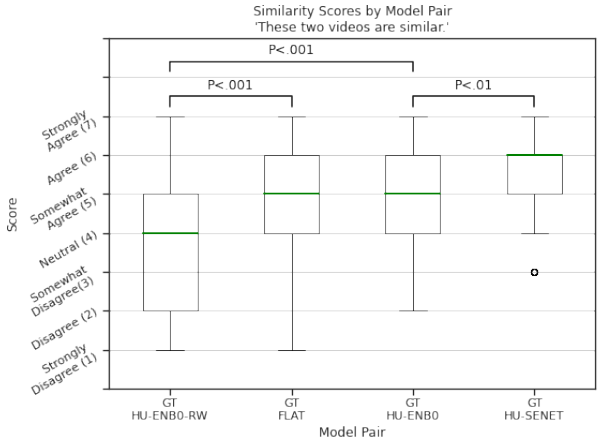

Fig. 4. Distribution of similarity scores by model.

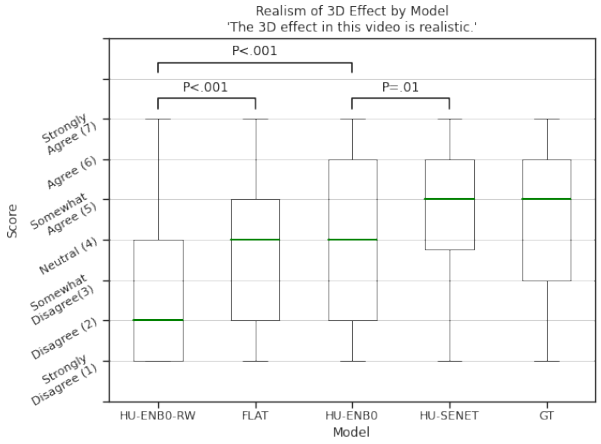

Fig. 5. Distribution of scores for realism of 3D effect by model.

(p=0.001). We also found a significant difference between HU-ENB0 and HU-SENET (p=0.013).

For the questions relating to the realism, the Kruskal-Wallis rank sum test indicated a main effect of MDE model on realism (Figure 5). Pairwise comparisons using Wilcoxon rank sum test (Bonferroni correction) indicated a significant difference between the model with randomised weights HU-ENB0-RW (MD=2) and all other models: FLAT (MD=4, $p < .001$), HU-ENB0 (MD=4, $p < .001$) and HU-SENET (MD=5, $p < .001$). We also measured significant differences between HU-ENB0 and HU-SENET (p=0.0138), but no significant difference between HU-SENET and GT (p=0.486).

D. Discussion

Overall, the results seem to support the standard benchmark metrics as both the user study and the image similarity benchmark agree on the general ranking of the models. We also see that SSIM and PSNR can struggle with geometric transforms and weaken the discriminative power of the image similarity benchmark, despite it being designed to mimic how a user may evaluate depth maps. It is interesting that no

significant difference was found between HU-ENB0 and the flat depth maps (i.e. no depth). Users only somewhat agreed that the outputs from HU-ENB0 and the flat depth maps were similar to the ground truth, and they neither agreed or disagreed that the 3D effect in these models' outputs are realistic. This would suggest that outputs from this model are not acceptable for use in VR. From this, we might infer that models with a relative error of around 0.141 are generally not accurate enough. Similarly, we could conclude that in order to be of acceptable accuracy for VR content creation, a model should have a relative error of around 0.115, at which point the output is considered to be of similar quality as ground truth data and provides a 3D effect that is about as realistic as ground truth data. Interestingly, 0.115 REL may be the point at which a model may perform better than data from a Kinect sensor. Users rated the 3D effect of outputs from HU-SENET higher more often than the ground truth (Figure 5) and there are certainly examples of the estimated depth appearing more realistic (Figure 1). However, we could not measure any significant differences.

E. Limitations and Future Work

It is important to mention that we only tested a limited range of depth estimation models that share a similar architecture and training regime. In future work, it will be important to compare a wider range of model architectures and model sizes to more accurately assess the threshold of what is could be considered 'good enough' for VR content creation. The scale of the user study was also limited as we only included four scenes from a single dataset. Future work should look at including a wider variety of scenes from different datasets to verify how well the results generalise. Improving the quality of the 3D meshes would help improve users' perceived quality of the rendered scenes, and provide an experience that would be closer to what one would expect from applications with 3D content (e.g. One-Shot 3D Photography [18]). It would also be of interest to run a similar study in a VR environment where users can move through the scenes themselves. This would likely change their perception of how realistic the 3D photographs appear. One final avenue for future work that comes to mind would be to compare whether the user study results would be consistent if the extracted depth data was used for 3D scene reconstruction.

VII. Conclusion

In this paper, we presented the first 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 different 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 with regard to creating 3D content. However we also found that some metrics, such as SSIM and PSNR are not always reliable enough to find difference that would be obvious to human observers. We hope that our work contributes to improving the quality assessments and benchmarking of depth estimation methods in particular with regard to VR content creation.

Acknowledgements

We gratefully acknowledge the support of the New Zealand Marsden Council through Grant UOO1724.

References

[1] S. Niklaus, L. Mai, J. Yang, and F. Liu, "3D Ken Burns Effect from a Single Image," *ACM Trans. Graph.*, vol. 38, no. 6, p. 184, 2019.

[2] P. Hedman and J. Kopf, "Instant 3d photography," *ACM Transactions on Graphics (TOG)*, vol. 37, no. 4, pp. 1–12, 2018.

[3] S. Zollmann, A. Dickson, and J. Ventura, "Casualvrvideos: Vr videos from casual stationary videos," in *26th ACM Symposium on Virtual Reality Software and Technology*, ser. VRST '20. New York, NY, USA: Association for Computing Machinery, 2020. [Online]. Available: https://doi.org/10.1145/3385956.3422119

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-Decem, 2016, pp. 770–778.

[5] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2002–2011.

[6] 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," *arXiv preprint arXiv:1907.10326*, 2019.

[7] K. Lasinger, R. Ranftl, K. Schindler, and V. Koltun, "Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-Shot Cross-Dataset Transfer," *arXiv preprint arXiv:1907.01341*, 2019.

[8] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," *arXiv preprint arXiv:2103.13413*, 2021.

[9] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from RGBD images," in *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 7576 LNCS, no. PART 5, 2012, pp. 746–760.

[10] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *Int. J. Rob. Res.*, vol. 32, no. 11, pp. 1231–1237, 2013.

[11] A. Dickson, A. Knott, and S. Zollmann, "User-centred Depth Estimation Benchmarking for VR Content Creation from Single Images," in *Pacific Graphics Short Papers, Posters, and Work-in-Progress Papers*, S.-h. Lee, S. Zollmann, M. Okabe, and B. Wuensche, Eds. The Eurographics Association, 2021, p. to appear.

[12] C. Cadena, Y. Latif, and I. D. Reid, "Measuring the performance of single image depth estimation methods," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2016, pp. 4150–4157.

[13] T. Koch, L. Liebel, F. Fraundorfer, and M. Korner, "Evaluation of cnn-based single-image depth estimation methods," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018, pp. 0–0.

[14] J. Hu, M. Ozay, Y. Zhang, and T. Okatani, "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.

[15] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, vol. 2. Ieee, 2003, pp. 1398–1402.

[16] A. Horé and D. Ziou, "Image quality metrics: Psnr vs. ssim," in *2010 20th International Conference on Pattern Recognition*, 2010, pp. 2366–2369.

[17] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "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.

[18] J. Kopf, K. Matzen, S. Alsisan, O. Quigley, F. Ge, Y. Chong, J. Patterson, J.-M. Frahm, S. Wu, M. Yu, P. Zhang, Z. He, P. Vajda, A. Saraf, and M. Cohen, "One shot 3d photography," *ACM Trans. Graph.*, vol. 39, no. 4, Jul. 2020.