

Performance evaluation of depth estimation on facial images

Matevž Rom¹, Manfred Gonzalez-Hernandez¹, Peter Peer¹, Vitomir Štruc², Blaž Meden¹

¹Faculty of Computer and Information Science, University of Ljubljana, Slovenia,

²Faculty of Electrical Engineering, University of Ljubljana, Slovenia

E-mail: matevz.rom@gmail.com

Abstract

The study of depth estimation from images leverages various datasets and models to enhance accuracy and quality. We analyze the CelebAMask-HQ dataset, known for its high-resolution celebrity images. Additionally, we analyze the Relative Human dataset, which offers diverse multi-person images with comprehensive annotations. We also investigate the C3I-SynFace dataset, notable for its large-scale synthetic human faces. Our evaluation encompasses the Intel Dense Prediction Transformer (DPT), TCMonoDepth, and the Depth Anything models. Employing histogram matching for depth image preprocessing, we compare models using general image quality metrics (MSE, MAE, SSIM, LPIPS) and overall visual analysis, particularly focusing on artifacts. Our results reveal that both Intel DPT and TCMonoDepth exhibit artifacts, whereas Depth Anything produces clearer depth images with fewer artifacts. Remarkably, Depth Anything surpasses other models in all metrics on the C3I-SynFace dataset, despite encountering challenges with noise and detail handling in specific scenarios.

1 Introduction

Depth estimation, the process of inferring the distance of objects from a camera or sensor in a scene, has been a significant area of research within computer vision for several decades. Traditional approaches to depth estimation include stereo vision and structured light techniques. Early stereo vision methods, such as the seminal work by Marr and Poggio (1976) [1], focused on reconstructing 3D scenes from multiple images taken from different viewpoints by matching corresponding points and calculating disparities. Similarly, structured light techniques, explored by Besl (1988) [2] and Krotkov (1987) [3], involved projecting a known pattern of light onto a scene and analyzing the deformation of the pattern to infer depth. These foundational methods have significantly influenced modern advancements in depth estimation and computer vision.

Recent years have seen the emergence of deep learning-based approaches for depth estimation, reducing the reliance on specialized sensors and enabling depth prediction from video samples without depth-based ground truth maps. These models have found applications across various domains, including robotics [4], 3D scene reconstruction

[5, 6] autonomous vehicles [7], and augmented reality [8], demonstrating their versatility and potential.

The focus of this research is to identify the most effective depth estimation model for future advancements in biometric applications, particularly within the domain of generative models. Depth information is crucial for ensuring the accurate and reliable performance of biometric data generation in real-world scenarios.

In this study, we analyze several depth estimation models and evaluate their performance on three datasets. The datasets include two collections of real-life human images and one of synthetic human faces. Our evaluation protocol involves measuring Mean Squared Error (MSE), Mean Absolute Error (MAE), Structural Similarity Index Measure (SSIM), Learned Perceptual Image Patch Similarity (LPIPS) and iterations per second (iter/sec), with a specific focus on the synthetic dataset to gauge the models' effectiveness in predicting depth from image-based biometric data.

2 Depth estimation models

Here we provide short description of the three state-of-the-art depth estimation models that we used in this study.

2.1 TCMonoDepth

The TCMonoDepth [9] method aims to enforce temporal consistency without the need for sequential modules or multi-frame processing.

The authors introduced a straightforward approach to improve temporal consistency by learning a prior from video data and incorporating it into existing single-image depth estimation methods. During testing, the method performs end-to-end inference frame by frame, achieving stable depth estimation while maintaining high quality.

Experimental results on benchmark datasets demonstrate superior temporal consistency compared to existing methods, without additional computational cost. Human subjective studies confirm the effectiveness of the proposed metric in assessing visual stability. Overall, the method offers a practical solution for video depth estimation in various applications, including augmented reality and autonomous driving.

2.2 Intel DPT

The Dense Prediction Transformer (DPT) model is a computer model that learns to estimate depth from images. It was trained using 1.4 million images and was first introduced in a research paper called “Vision Transformers for Dense Prediction” [10]. The model uses a type of neural network called a Vision Transformer (ViT) [11] as its main structure and then adds additional layers on top to help it estimate depth from single images. The Intel DPT model is publicly available.

2.3 Depth Anything

Depth Anything model [12] was trained on more than 62 million unlabeled images. Authors did the training by designing automatic annotation and large scale collection of unlabeled visual data for enlarging coverage and reducing generalization error. This model ended up being significantly better than the previously best relative MDE model MiDaS v3.1 [13].

3 Evaluation datasets

In the following section, we briefly describe each dataset that was used for evaluation.

3.1 CelebAMask-HQ

CelebAMask-HQ [14] is a large-scale face image dataset containing 30,000 high-resolution images selected from the CelebA dataset, following the CelebA-HQ standard. Each image is annotated with segmentation masks for 19 facial attributes, including all facial components and accessories such as skin, nose, eyes, eyebrows, ears, mouth, lips, hair, hat, eyeglasses, earrings, necklaces, neck, and clothing. These masks were manually annotated and correspond to the CelebA dataset. CelebAMask-HQ can be used to train and evaluate algorithms for face parsing, face recognition, and GANs for face generation and editing.

3.2 Relative Human

Relative Human dataset contains multi-person in-the-wild RGB images with extensive human annotations, as described in [15]. The annotations include relative depth relationships and ordering between all individuals in the image, age group classifications (adults, teenagers, kids, babies), gender, bounding boxes, and 2D poses. The dataset spans a diverse range of age groups from babies to elderly adults, featuring various ethnicity labels, and includes real-life backgrounds such as bars, churches, living rooms, beaches, and more.

3.3 C3I-SynFace

C3I-SynFace includes large-scale synthetic human face dataset with its corresponding ground truth images [16]. The dataset is created using 15 female and 15 male synthetic 3D human models from iClone software in FBX format. Five facial expressions—neutral, angry, sad, happy, and scared—are incorporated to introduce further variations. Utilizing these models, an open-source data generation pipeline in Python is developed to import them into

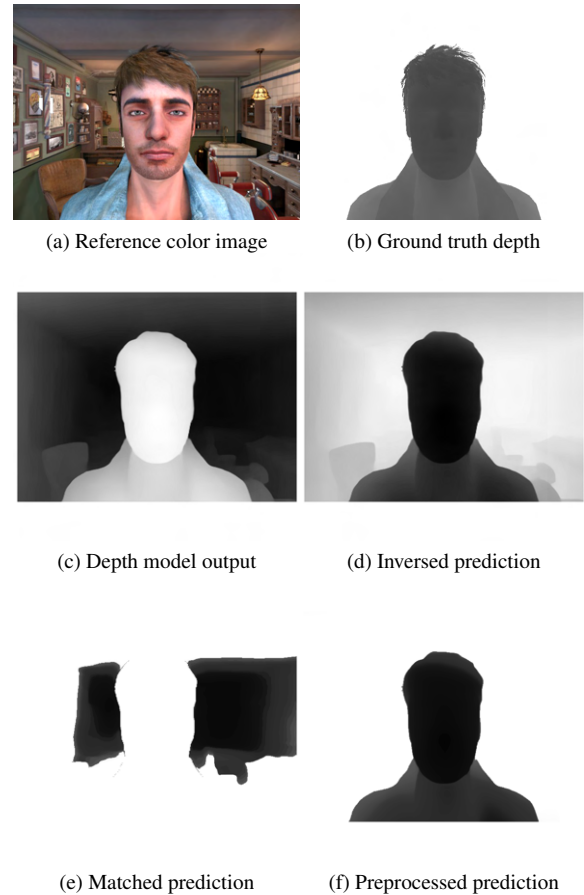


Figure 1: Reference color image and depth image from C3I-SynFace [16] and images, generated with Depth Anything [12] during preprocessing: with and without inversion, and with and without histogram matching.

Blender, a 3D computer graphics tool, and render facial images with ground truth annotations of head pose and face depth in raw format. This framework, leveraging virtual human models, can generate extensive synthetic facial datasets (e.g., head pose or face depth datasets) with high control over facial and environmental variations such as pose, illumination, and background.

4 Results

Our evaluation protocol included prediction of depth maps on selected evaluation datasets and analysis of the obtained results. Prior to model evaluation it was necessary to perform histogram matching [17] as a preprocessing step on predicted depth maps.

4.1 Histogram matching

Before computing the MSE, depth images had to be preprocessed. The gray scale images were inverted and their histograms were matched with ground truth images. The reason for inverting the images is because the ground truth uses the opposite ends for closeness and distances compared to depth models and without this the histogram matching produces unwanted results. This can be observed on Figures 1c, 1d, 1e, 1f.

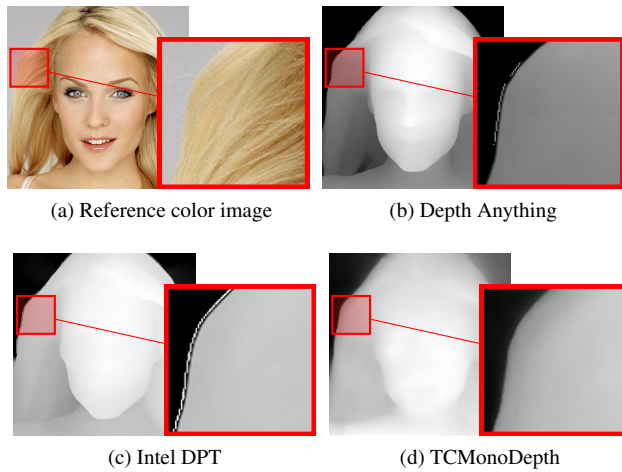


Figure 2: Artifacts comparison between the models. Although Depth Anything provides best overall depth prediction, it sometimes introduces noisy artefacts concentrated on intensive edge changes. Similarly, Intel DPT also produces these artefacts, which are even more visible. TCMonoDepth does not have this issue, although the details of its predicted depth maps is not as high as with other two models.

4.2 Model performance comparison

In this section, we present different aspects of our performance analysis of depth estimation models, including the qualitative (visual artefacts analysis, reconstructed depth maps comparison, preserved high frequency details comparison) and quantitative experiments (reconstruction error analysis is performed by computing MSE, MAE, SSIM, and LPIPS on synthetic facial images, comparing ground truth depth masks with reconstructed depth masks)

4.2.1 Qualitative performance analysis

Both Intel DPT and Depth Anything ended up producing artefacts. Contrary, TCMonoDepth didn't produced as many on our evaluation datasets. The resulting artifacts are illustrated in Figure 2.

Among the three models, Intel DPT ended up producing the blurriest images. Results of TCMonoDepth and Depth Anything are more detailed in favor of Depth Anything. See Figure 3. The problems seem to be amplified on close up shots like in CelebaMask-HQ dataset and isn't as evident on pictures taken from afar since the blurriness is similar between the models. Some depth images produced noise and as we can see on Figure 3, while Depth Anything seems to produce the least amount of it and gives visually better results.

4.2.2 Quantitative performance analysis

Despite the fact that images had more line artifacts with Depth Anything, all the scores measured on synthetic images seem to be better compared to other models (Table 1). We can assume that the reason for Depth Anything outperforming TCMonoDepth model is that synthetic images had less noise so Intel DPT model doesn't have that much blurry results compared to datasets of real



Figure 3: Visual comparison of predicted results on CelebAMask-HQ (row 1 to 4) and Relative Human (row 5 to 8) datasets. From left to right: Source RGB image, followed by depth maps estimated with TCMonoDepth, Intel DPT, and Depth Anything, respectively.

images. With SSIM score we can observe that all models produced images very similar to ground truth with Depth Anything being slightly better than the rest.

Based on the results obtained it can be seen that Depth Anything achieved best performance in all metrics, indicating the closest resemblance to ground truth depth maps, highlighting that this model is able to estimate most accurate depth information amongst the three techniques. The reason for such a big difference in mean value of MSE between TCMonoDepth and the others can be seen in Figure 4. TCMonoDepth picked up much more background noise and with it comes increased error in MSE. Intel DPT model performs somewhere in between the two techniques.

The primary challenge for depth estimation algorithms was handling image content with a high level of detail, such as hair. Images with intricate details, particularly

Table 1: Mean Squared Error (MSE), Mean Absolute Error (MAE), Structural Similarity Index (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS) on C3I-SynFace.

	Intel DPT	Depth Anything	TCMonoDepth
MSE↓	10.62 ± 3.38	9.65 ± 2.67	18.75 ± 4.36
MAE↓	37.68 ± 6.11	36.47 ± 3.72	44.39 ± 5.38
SSIM↑	0.95 ± 0.017	0.96 ± 0.01	0.92 ± 0.019
LPIPS↓	0.092 ± 0.036	0.064 ± 0.02	0.139 ± 0.04
iter/sec↑	10.5	2.8	19.2



Figure 4: Reference color image (left), ground truth depth image (middle), and induced prediction noise (right), produced by TCMonoDepth (on C3I-SynFace dataset).

those containing high-frequency elements like pointed hair, generated a higher number of gray pixels in depth estimation compared to synthetic images, as shown in Figure 5. This issue may be less pronounced in real-world photographs, where obtaining ground truth data for each individual strand of hair is more difficult than for synthetic images.



Figure 5: Side by side difference in high frequency image details (especially hair) between ground truth (middle) and Depth Anything prediction (right) on C3I-SynFace.

5 Conclusion

In this preliminary study on the general capability of depth estimation models, we analyzed a few (real-world and synthetic) datasets: CelebAMask-HQ, Relative Human and C3I-SynFace. We evaluated the performance of three depth estimation models: Intel Dense Prediction Transformer (Intel DPT), TCMonoDepth, and Depth Anything.

Our analysis revealed three key findings, namely: (a) The level of artifact generation: Both the Intel DPT and Depth Anything models produced artifacts, such as white line edges around objects, while the TCMonoDepth model did not produce any artifacts on the datasets tested, (b) Depth map clarity / induced noise: The Depth Anything model generated the clearest depth images with the least amount of noise. However, the TCMonoDepth generated the blurriest images with significant amount of background noise, which negatively impacted its performance, and (c) Depth Anything outperformed all other models in synthetic dataset-based performance metrics. In conclusion, Depth Anything excels in producing clearer and

more visually appealing depth images with fewer artifacts than Intel DPT model and shows superior performance in all tested performance metrics on synthetic dataset. TCMonoDepth, though free from artifacts, suffers from significant noise issues. The findings highlight the trade-off between image clarity, artifact generation, and error metrics in depth estimation models. Our future work will focus on applying analyzed models into the development of privacy-preserving biometrics [18].

References

- [1] D. Marr and T. Poggio, "Cooperative computation of stereo disparity," *Science*, vol. 194, no. 4262, pp. 283–287, 1976.
- [2] P. J. Besl, "Active, optical range imaging sensors," *Machine Vision and Applications*, vol. 1, no. 2, pp. 127–152, 1988.
- [3] E. P. Krotkov, "Focusing," *International Journal of Computer Vision*, vol. 1, no. 3, pp. 223–237, 1987.
- [4] M. Ye, E. Johns, A. Handa, L. Zhang, P. Pratt, and G.-Z. Yang, "Self-supervised siamese learning on stereo image pairs for depth estimation in robotic surgery," 2017.
- [5] T. Oblak, J. Šircelj, V. Štruc, P. Peer, F. Solina, and A. Jaklič, "Learning to predict superquadric parameters from depth images with explicit and implicit supervision," *IEEE Access*, vol. 9, pp. 1087–1102, 2021.
- [6] D. Tomašević, P. Peer, F. Solina, A. Jaklič, and V. Štruc, "Reconstructing superquadrics from intensity and color images," *MDPI Sensors*, vol. 22, no. 14, p. 5332, 2022.
- [7] M. Fonder, D. Ernst, and M. V. Droogenbroeck, "M4depth: Monocular depth estimation for autonomous vehicles in unseen environments," 2022.
- [8] M. Kim, S. H. Choi, K.-B. Park, and J. Y. Lee, "A hybrid approach to industrial augmented reality using deep learning-based facility segmentation and depth prediction," *Sensors*, vol. 21, no. 1, 2021.
- [9] S. Li, Y. Luo, Y. Zhu, X. Zhao, Y. Li, and Y. Shan, "Enforcing temporal consistency in video depth estimation," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2021.
- [10] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," 2021.
- [11] 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," 2021.
- [12] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. Zhao, "Depth anything: Unleashing the power of large-scale unlabeled data," in *CVPR*, 2024.
- [13] R. Birkel, D. Wofk, and M. Müller, "Midas v3.1 – a model zoo for robust monocular relative depth estimation," 2023.
- [14] C.-H. Lee, Z. Liu, L. Wu, and P. Luo, "Maskgan: Towards diverse and interactive facial image manipulation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [15] Y. Sun, W. Liu, Q. Bao, Y. Fu, T. Mei, and M. J. Black, "Putting people in their place: Monocular regression of 3D people in depth," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [16] S. Basak, F. Khan, H. Javidnia, R. M. Peter Corcoran, and M. Schukat, "C3i-synface: A synthetic head pose and facial depth dataset using seed virtual human models," in *Data in Brief*, 2023.
- [17] R. C. Gonzalez, *Digital image processing*. Pearson education india, 2009.
- [18] B. Meden, P. Rot, P. Terhörst, N. Damer, A. Kuijper, W. J. Scheirer, A. Ross, P. Peer, and V. Štruc, "Privacy-enhancing face biometrics: A comprehensive survey," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 4147–4183, 2021.