Contents lists available at ScienceDirect

Computers & Graphics

journal homepage: www.elsevier.com/locate/cag

# Vision UFormer: Long-Range Monocular Absolute Depth Estimation Supplementary Materials

Tomas Polasek[a], Martin Čadík[a,*], Yosi Keller[b], Bedrich Benes[c]

[a] *Faculty of Information Technology, Brno University of Technology, Bozetechova 2/1, Brno, 612 00, Czech Republic*
[b] *Faculty of Engineering, Bar-Ilan University, Ramat Gan, 1102-1105, Israel*
[c] *305 N University St., Purdue University, West Lafayette, IN 47907-2107, United States of America*

ARTICLE INFO

*Article history*:
Received February 20, 2023

*2000 MSC:* 94A08, 68U10

*Keywords:* Absolute Depth Prediction, Monocular, Long-range, Transformer

ABSTRACT

We introduce Vision UFormer (ViUT), a novel deep neural long-range monocular depth estimator. The input is an RGB image, and the output is an image that stores the absolute distance of the object in the scene as its per-pixel values. ViUT consists of a Transformer encoder and a ResNet decoder combined with the UNet style of skip connections. It is trained on 1M images across ten datasets in a staged regime that starts with easier-to-predict data such as indoor photographs and continues to more complex long-range outdoor scenes. We show that ViUT provides comparable results for normalized relative distances and short-range classical datasets such as NYUv2 and KITTI. We further show that it successfully estimates absolute long-range depth in meters. We validate ViUT on a wide variety of long-range scenes showing its high estimation capabilities with a relative improvement of up to 23%. Absolute depth estimation finds application in many areas, and we show its usability in image composition, range annotation, defocus, and scene reconstruction. Our models are available at **cphoto.fit.vutbr.cz/viut**.

## 1. Introduction

This document contains supplementary materials for the "Vision UFormer: Long-Range Monocular Absolute Depth Estimation" paper.

## 2. Proposed Method

### *2.1. Training*

During training, we utilize multiple modalities of data depending on their availability within the current training dataset. In order to train on the other modalities, we use the following additional loss functions along with the primary $\mathcal{L}_{si}$:

- Segmentation: Cross-Entropy Loss
- Instances: Cross-Entropy Loss
- Normals: Mean Squared Error Loss
- Optical Flow: L2 Loss
- Diffuse Color: Mean Squared Error Loss
- Shading: Mean Squared Error Loss

During the training, we alternate between batches of depth and additional modalities. Each time we switch the model's heads to the corresponding modality in order to prevent changing weights for the heads of other modalities.

### *2.2. Datasets*

We perform pre-processing on each of the ten training datasets to improve their usability in the staged training.

[*] Corresponding author: Tel.: +420 54114 1272
*e-mail:* ipolasek@fit.vut.cz (Tomas Polasek), cadik@fit.vut.cz (Martin Čadík), yosi.keller@gmail.com (Yosi Keller), bbenes@purdue.edu (Bedrich Benes)

| | Test → | GP3K [1] | | | | | | | LSAR [2] | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | Train ↓ | RMS | REL | Log10 | $\delta >$ | $1.25^1$ | $1.25^2$ | $1.25^3$ | RMS | REL | Log10 | $\delta >$ | $1.25^1$ | $1.25^2$ | $1.25^3$ |
| ViUT (**ours**) | Staged | 128.875 | 0.104 | 0.073 | | 13.91 | 2.18 | 0.31 | 142.079 | 0.113 | 0.078 | | 14.84 | 2.62 | 0.52 |
| DPT [3] | MIX 6 [3] | 168.569 | 0.134 | 0.090 | | 17.87 | 4.33 | 1.27 | 184.484 | 0.140 | 0.094 | | 19.48 | 5.45 | 1.82 |
| AdaBins [4] | KITTI [5] | 184.471 | 0.144 | 0.097 | | 18.47 | 4.74 | 1.47 | 188.189 | 0.149 | 0.102 | | 19.94 | 5.83 | 1.99 |
| MiDaS [6] | MIX 5 [3] | 254.347 | 0.212 | 0.150 | | 24.82 | 10.46 | 4.32 | 269.152 | 0.218 | 0.156 | | 26.48 | 12.40 | 5.41 |
| MegaDepth [7] | MegaDepth [7] | 933.634 | 0.739 | 0.343 | | 37.47 | 27.42 | 20.67 | 950.427 | 0.746 | 0.344 | | 39.74 | 29.90 | 22.79 |
| Pix2Pix [8] | Mannequin [8] | 1626.690 | 1.363 | 0.390 | | 43.27 | 31.70 | 25.35 | 1646.842 | 1.370 | 0.391 | | 43.24 | 31.68 | 25.33 |
| WSVD [9] | WSVD [9] | 903.431 | 0.714 | 0.343 | | 32.86 | 21.84 | 13.16 | 919.772 | 0.720 | 0.343 | | 34.74 | 24.88 | 17.04 |

Table 1: **Depth Estimation:** Additional evaluation results of various state-of-the-art techniques on the Geopose3K [1] and LandscapeAR [2] datasets. For a definition of the metrics, see Sec. 4.2 of the main paper.

**EDEN [10]** We train on the raw depths, taking their values as absolute distances in meters. The other modalities are trained by using the specific loss functions as specified above. For optical flow, we use the *ForwardFlow* files. For shading, we use the *Glossy* color outputs.

**SINTEL [11]** We use the final render, including all of the effects as the input. We scale the depths using min-max scaling to reduce them from the original range into 0.0 to 1.0 values. We use 100m as the maximum, clipping any values that go outside of this range. Finally, we use the clean renders without effects as the prediction target for the Shading modality.

**DIW [12]** We pre-process the ordinal point-wise depth relations into sparse relative depth maps. We then train on these as two separate modalities. We use Binary Cross-Entropy Loss for the point-wise relations and Mean Squared Error for the sparse maps. During training, we use each image exactly once, even when it is assigned with multiple user inputs. We also randomly sample the images into input batches so that each training epoch does not contain all available training inputs.

**NYU [13]** We use the original depths without scaling and utilize the provided instances and segmentation maps.

**TUM [14]** We automatically align the RGB and depth sequences by searching for the closest match. For this, we use the provided timestamps as suggested by the authors. We use the depths without any additional scaling.

**MegaDepth [7]** For training, we use both the relative and ordinal depth maps as two separate modalities. We clip the relative depths to 200 units and linearly scale them into a 0 to 1 range.

**ETH3D [15]** We clip the depths to 50m and otherwise use them as provided by the authors.

**KITTI [5]** We only use the subset of the KITTI dataset with registered depths, without any additional modalities. We clip the distances to 20km and follow by linearly scaling them into a 0 to 1 interval.

**Geopose3K [1]** We use both the provided renders and photos as separate input modalities. Since the depths are synthetic and miss many of the details, we generate automatically generate masks and train only the regions covered purely by the terrain. The masks are generated by using a combination of DeepLabv3 [16] segmentation along with further processing and filtering to ensure that only reliable data is used for training. We also use the negative depth values, which signify the sky, to enhance our training masks. To better work with the large ranges of depth, we use non-linear scaling, similar to a z-buffer:

$$z(d) = \frac{far near}{near + d(far - near)},$$

where $d$ is the linear depth, $z$ is the non-linear logarithmic depth, and we set $near = 0$ and $far = 250$km to cover the values present in the dataset. Depth values outside of this range are clipped.

**LandscapeAR [2]** We perform similar processing as in the case of Geopose3K. We choose a smaller subset of valid segmentation classes to ensure the quality of the training data. For the depth, we again use the non-linear scaling as presented above, setting $near = 0$ and $far = 300$km to better cover this dataset's increased range.

## 3. Implementation, Experiments and Results

We base our quantitative analysis of the performance on the following commonly accepted evaluation metrics [17, 18, 19]:

$$\text{RMS} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left\|d_i - d'_i\right\|^2}, \tag{1}$$

$$\text{REL} = \frac{1}{N}\sum_{i=1}^{N}\frac{\left|d_i - d'_i\right|}{d_i}, \tag{2}$$

$$\text{Log10} = \frac{1}{N}\sum_{i=1}^{N}\left|log_{10}(d_i) - log_{10}(d'_i)\right|, \tag{3}$$

$$\delta > thr : \%\text{ of } d_i \text{ such that } \max\left(\frac{d_i}{d'_i}, \frac{d'_i}{d_i}\right) > thr, \tag{4}$$

$$\text{WHDR} = \frac{1}{\sum \omega_{ij}}\sum_{ij}^{A}\omega_{ij}\mathbf{1}(\ell_{ij} \neq \bar{\ell}_{ij}), \tag{5}$$

where $d_i$ is the ground-truth depth, $d'_i$ is the predicted depth, and $N$ is the size of the test set. For the WHDR [18], we set the weights $\omega_{ij}$ to 1 for all annotations $\ell_{ij}$.

We provide additional metrics calculated for the Geopose3K [1] and LandscapeAR [2] datasets in Tab. 1. Additional example pairs of input RGB and depths predicted by the `ViUT` can be seen in Fig. 1.

## References

[1] Brejcha, J, Čadík, M. Geopose3k: Mountain landscape dataset for camera pose estimation in outdoor environments. Image and Vision Computing 2017;66:1–14. URL: https://www.sciencedirect.com/science/article/pii/S0262885617300963. doi:https://doi.org/10.1016/j.imavis.2017.05.009.

[2] Brejcha, J, Lukac, M, Hold-Geoffroy, Y, Wang, O, Cadik, M. Landscapear: Large scale outdoor augmented reality by matching photographs with terrain models using learned descriptors. In: European Conference on Computer Vision (ECCV). Springer International Publishing. ISBN 978-3-030-58526-6; 2020, p. 295–312.

[3] Ranftl, R, Bochkovskiy, A, Koltun, V. Vision transformers for dense prediction. In: IEEE International Conference on Computer Vision (ICCV). 2021, p. 12179–12188.

[4] Bhat, SF, Alhashim, I, Wonka, P. Adabins: Depth estimation using adaptive bins. In: IEEE Computer Vision and Pattern Recognition (CVPR). 2021, p. 4009–4018.

[5] Geiger, A, Lenz, P, Urtasun, R. Are we ready for autonomous driving? the kitti vision benchmark suite. In: IEEE Computer Vision and Pattern Recognition (CVPR). 2012, p. 3354–3361. doi:10.1109/CVPR.2012.6248074.

[6] Ranftl, R, Lasinger, K, Hafner, D, Schindler, K, Koltun, V. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE Trans Pattern Anal Mach Intell 2022;44(3):1623–1637. doi:10.1109/TPAMI.2020.3019967.

[7] Li, Z, Snavely, N. Megadepth: Learning single-view depth prediction from internet photos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018,.

[8] Li, Z, Dekel, T, Cole, F, Tucker, R, Snavely, N, Liu, C, et al. Learning the depths of moving people by watching frozen people. In: IEEE Computer Vision and Pattern Recognition (CVPR). 2019,.

[9] Wang, C, Lucey, S, Perazzi, F, Wang, O. Web stereo video supervision for depth prediction from dynamic scenes. In: 2019 International Conference on 3D Vision (3DV). 2019, p. 348–357. doi:10.1109/3DV.2019.00046.

[10] Le, HA, Mensink, T, Das, P, Karaoglu, S, Gevers, T. Eden: Multimodal synthetic dataset of enclosed garden scenes. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). 2021, p. 1579–1589.

[11] Butler, DJ, Wulff, J, Stanley, GB, Black, MJ. A naturalistic open source movie for optical flow evaluation. In: European Conference on Computer Vision (ECCV). Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN 978-3-642-33783-3; 2012, p. 611–625.

[12] Chen, W, Fu, Z, Yang, D, Deng, J. Single-image depth perception in the wild. In: Lee, D, Sugiyama, M, Luxburg, U, Guyon, I, Garnett, R, editors. Advances in Neural Information Processing Systems; vol. 29. Curran Associates, Inc.; 2016,URL: https://proceedings.neurips.cc/paper/2016/file/0deb1c54814305ca9ad266f53bc82511-Paper.pdf.

[13] Silberman, N, Hoiem, D, Kohli, P, Fergus, R. Indoor segmentation and support inference from rgbd images. In: European Conference on Computer Vision (ECCV). Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN 978-3-642-33715-4; 2012, p. 746–760.

[14] Sturm, J, Engelhard, N, Endres, F, Burgard, W, Cremers, D. A benchmark for the evaluation of rgb-d slam systems. In: Proc. of the International Conference on Intelligent Robot Systems (IROS). 2012,.

[15] Schops, T, Schonberger, JL, Galliani, S, Sattler, T, Schindler, K, Pollefeys, M, et al. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In: IEEE Computer Vision and Pattern Recognition (CVPR). 2017,.

[16] Chen, LC, Zhu, Y, Papandreou, G, Schroff, F, Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: ECCV. 2018,.

[17] Eigen, D, Puhrsch, C, Fergus, R. Depth map prediction from a single image using a multi-scale deep network. In: Adv. in Neural Inf. Proc. Systems (NIPS); vol. 27. Curran Associates, Inc.; 2014,URL: https://proceedings.neurips.cc/paper/2014/file/7bccfde7714a1ebadf06c5f4cea752c1-Paper.pdf.

[18] Zoran, D, Isola, P, Krishnan, D, Freeman, WT. Learning ordinal relationships for mid-level vision. In: IEEE International Conference on Computer Vision (ICCV). 2015,.

[19] Xian, K, Shen, C, Cao, Z, Lu, H, Xiao, Y, Li, R, et al. Monocular relative depth perception with web stereo data supervision. In: IEEE Computer Vision and Pattern Recognition (CVPR). 2018,.

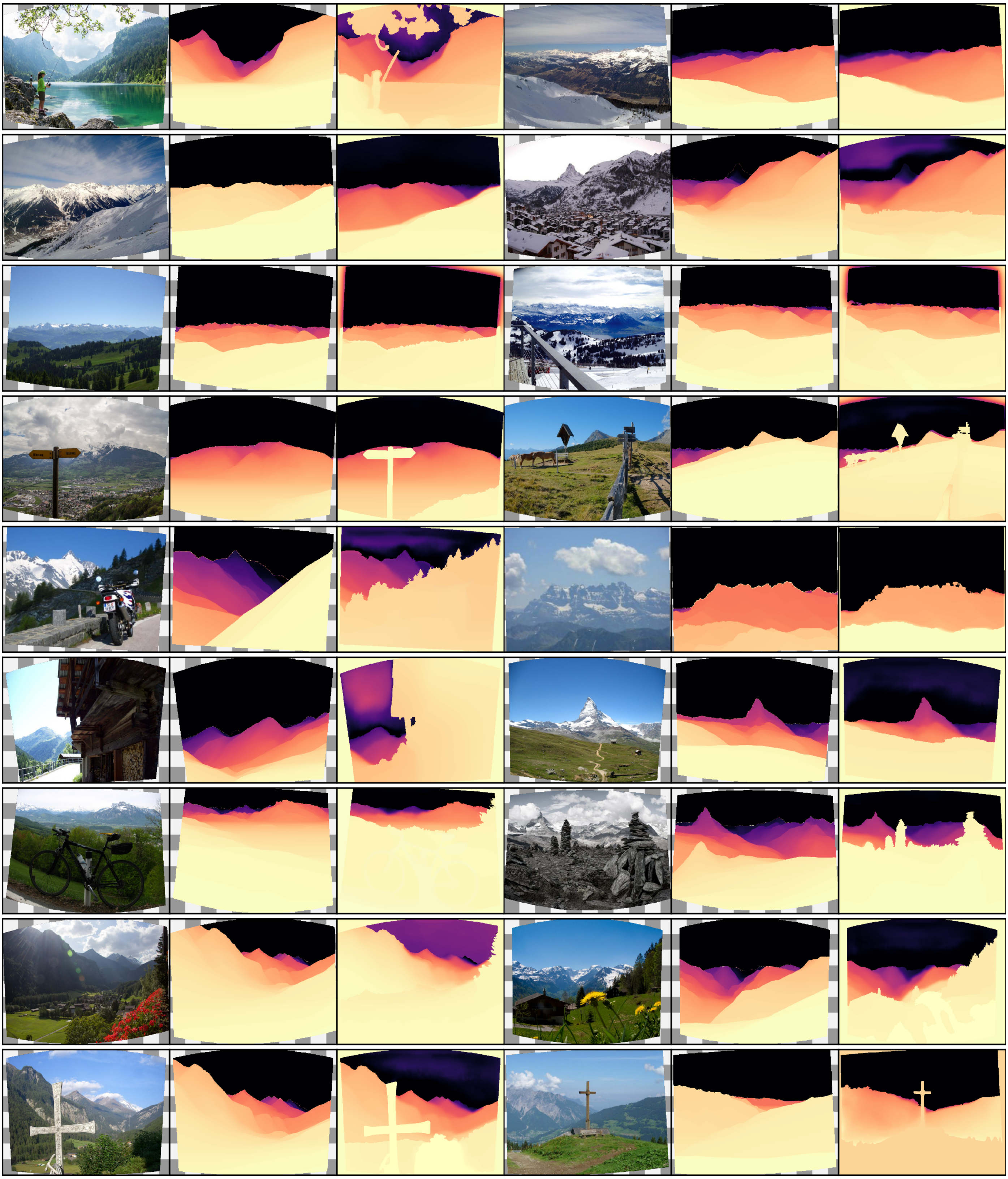

Fig. 1: **Depth Predictions:** Examples of depth maps predicted by the `ViUT` model. The images contain the input RGB image, visualized ground-truth depth, and a visualization of the predicted depth.