Pacific Graphics (2023)
R. Chaine, Z. Deng, and M. H. Kim (Editors)

Poster

Multi-scale Monocular Panorama Depth Estimation

Payal Mohadikar[1], Chuanmao Fan[1], Chenxi Zhao[1], and Ye Duan[2]

[1]University of Missouri, Missouri, USA
[2]Clemson University, South Carolina, USA

Abstract

Panorama images are widely used for scene depth estimation as they provide comprehensive scene representation. The existing deep-learning monocular panorama depth estimation networks produce inconsistent, discontinuous, and poor-quality depth maps. To overcome this, we propose a novel multi-scale monocular panorama depth estimation framework. We use a coarse-to-fine depth estimation approach, where multi-scale tangent perspective images, projected from 360 images, are given to coarse and fine encoder-decoder networks to produce multi-scale perspective depth maps, that are merged to get low and high-resolution 360 depth maps. The coarse branch extracts holistic features that guide fine branch extracted features using a Multi-Scale Feature Fusion (MSFF) module at the network bottleneck. The performed experiments on the Stanford2D3D benchmark dataset show that our model outperforms the existing methods, producing consistent, smooth, structure-detailed, and accurate depth maps.

CCS Concepts

• ***Computing methodologies*** → ***Scene understanding;***

1. Introduction and Methodology

Scene understanding has been a widely researched topic. Currently, models use a single EquiRectangular Projection (ERP) image to estimate depth. However, ERP image shows spherical distortion due to which the existing models try to solve this issue but still produce inconsistent, merging artifacts, and poor-quality depth maps. We therefore propose a novel multi-scale monocular panorama depth estimation model. Our model has a coarse and fine branch as shown in Figure 1. The coarse branch focuses on learning more holistic context information using low-resolution ERP images as input, while the fine branch uses high-resolution ERP images. We project input ERP images into multiple perspective tangent images for both branches using gnomonic projection. These perspective images are given to the encoder-decoder network to produce perspective depth maps which are merged to get final ERP low and high-resolution depth maps at output. Our coarse encoder takes coarse tangent patches and their 3D geometric embedding as input. To produce 3D geometric position embedding we use a Multi-Layer Perceptron (MLP) network that takes tangent pixel spherical co-ordinates (λ, ϕ, ρ) on the unit sphere and patch center co-ordinates (λ', ϕ') as input similar to OmniFusion [LGY*22]. The encoder output is then given to the Multi-head Self-Attention (MSA) module to capture long-range dependencies of tangent patch features. Like the coarse branch, we get fine-level bottleneck features from fine branch encoder-MSA network. To overcome fine-level depth map discrepancies that miss holistic contextual information, we use Multi-Scale Feature Fusion (MSFF) module. This module takes coarse and fine-level bottleneck features and produces their attention maps using convolution layers. These learned attention maps are multiplied with their feature maps and aggregated to produce updated fine-level features. The updated fine-level feature now has necessary holistic information. The coarse and updated fine-level features at the bottleneck are then given to coarse and fine decoder networks respectively which consist of up-sampling layers, convolution layers, and skip connections from encoder network. Finally, at the output of the decoders, we get low and high-resolution depth maps. Similar to OmniFusion [LGY*22] we refine the estimated depths iteratively by using depth values from previous iteration to update geometric information for MLP networks of next iteration. We train our network in an end-to-end manner and use BerHu loss [LRB*16] for supervision. The final loss function is addition of coarse and fine-depth BerHu loss, summed over all the iterations.

2. Experiments and Results

We evaluated our model performance using the Stanford2D3D [ASZS17] benchmark dataset. Table 1. shows comparative quantitative results using commonly used metrics. Figure 2. shows comparative qualitative results. Our model produces depth maps with no local patch merging artifacts, more accuracy, smoothness, structure details, and sharper boundaries than OmniFusion [LGY*22].

3. Acknowledgements

This research was partially supported by the National Science Foundation under award CNS-2018850, National Institute of

DOI: 10.2312/pg.20231282

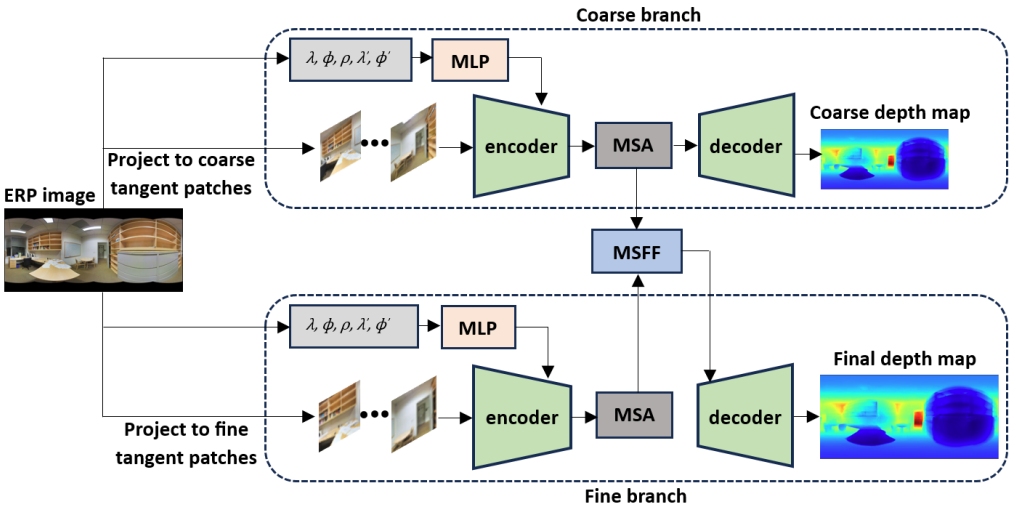

Figure 1: *Overall pipeline of Multi-scale Monocular Panorama Depth Estimation model.*

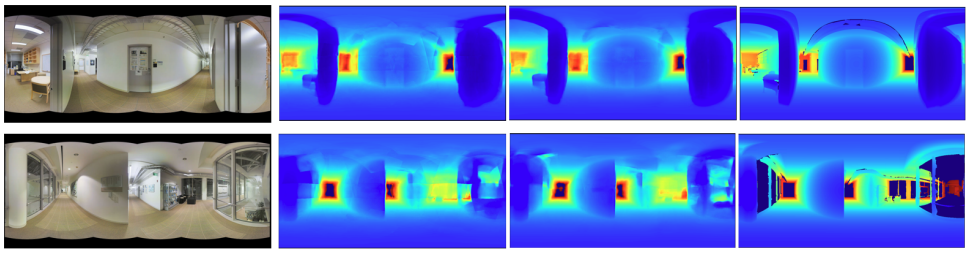

Figure 2: *Qualitative results on Stanford2D3D [ASZS17] benchmark dataset, ERP image (first column), OmniFusion [LGY*22] results (second column), Our model results (third column), and Ground truth (fourth column).*

Methods	Abs Rel↓	RMSE↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
FCRN [LRB*16]	0.1837	0.5774	0.7230	0.9207	0.9731
BiFuse [WYS*20]	0.1209	0.4142	0.8660	0.9580	0.9860
UniFuse [JSZ*21]	0.1114	0.3691	0.8711	0.9664	0.9882
HoHoNet [SSC21]	0.1014	0.3834	0.9054	0.9693	0.9886
OmniFusion [LGY*22]	0.0943	0.3582	0.8999	0.9742	0.9914
Ours	**0.0895**	**0.3423**	**0.9112**	**0.9759**	**0.9921**

Table 1: *Quantitative results on Stanford2D3D [ASZS17] benchmark dataset. Our model outperforms all the existing models for all metrics.*

Health under awards NIBIB-R01-EB02943, and U.S. Army Research Laboratory W911NF2120275. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the U. S. Government or agency thereof.

References

[ASZS17] ARMENI I., SAX S., ZAMIR A. R., SAVARESE S.: Joint 2d-3d-semantic data for indoor scene understanding. *arXiv preprint arXiv:1702.01105* (2017). 1, 2

[JSZ*21] JIANG H., SHENG Z., ZHU S., DONG Z., HUANG R.: Unifuse: Unidirectional fusion for 360 panorama depth estimation. *IEEE Robotics and Automation Letters 6*, 2 (2021), 1519–1526. 2

[LGY*22] LI Y., GUO Y., YAN Z., HUANG X., DUAN Y., REN L.: Omnifusion: 360 monocular depth estimation via geometry-aware fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 2801–2810. 1, 2

[LRB*16] LAINA I., RUPPRECHT C., BELAGIANNIS V., TOMBARI F., NAVAB N.: Deeper depth prediction with fully convolutional residual networks. In *2016 Fourth international conference on 3D vision (3DV)* (2016), IEEE, pp. 239–248. 1, 2

[SSC21] SUN C., SUN M., CHEN H.-T.: Hohonet: 360 indoor holistic understanding with latent horizontal features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 2573–2582. 2

[WYS*20] WANG F.-E., YEH Y.-H., SUN M., CHIU W.-C., TSAI Y.-H.: Bifuse: Monocular 360 depth estimation via bi-projection fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 462–471. 2