PG23

Daejeon, Korea

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

PROBLEM

Monocular 360-degree images are being popularly used for comprehensive scene understanding due to their integrated field of view. Existing monocular 360-degree image deep learning depth estimation networks produce inconsistent depth maps that miss fine structure details. To address this problem, we propose a novel multi-scale monocular panorama depth estimation framework to produce consistent, smooth, and accurate depth maps.

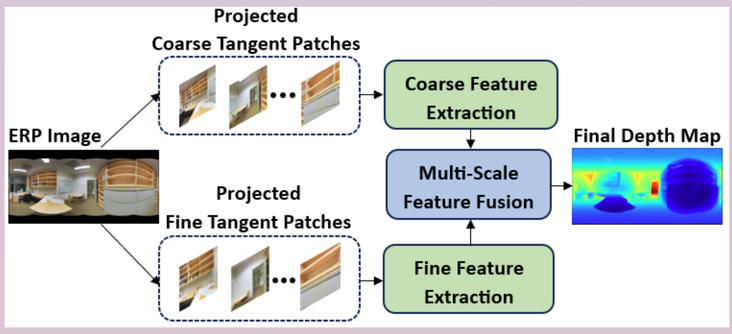

Figure 1. Multi-scale Monocular 360 Depth Estimation Network Overview

RELATED WORK

Recently, many models focus on using monocular EquiRectangular Projection (ERP) images to produce depth maps due to their easy availability, feasibility and integrated scene representation. ERP images however show distortions in polar regions making it challenging for existing distortion–free designed neural networks to perform depth estimation task. To address this issue existing models use distortion aware, spherical kernel etc. convolution neural networks [1]. Some networks [2, 3] use cube projection and ERP aggregated features. Recent networks show improvement by using horizontal latent representation [4] or projected tangent images to predict perspective depth maps which are merged to get ERP depth maps [5]. However, these local tangent patches predicted depth maps miss holistic geometric context information leading to inconsistent, patch merging artifacts, and poor quality depth maps.

OVERVIEW

We propose a novel multi-scale monocular 360 image depth estimation network, whose overview is shown in Figure 1. Our pipeline has coarse and fine branch that takes low and high-resolution ERP images. These ERP images are projected to multi-scale tangent patches. These patches then along with their 3D geometric embedded information are given to the encoder-decoder network to produce low and high-resolution tangent perspective depth maps which are merged to get final coarse and fine ERP depth maps. To overcome the fine depth map inconsistency due to the missing global context we propose a Multi-Scale Feature Fusion (MSFF) module that learns to guide the fine level image features with global contextual information using attention fusion at the network bottleneck. Our model outperforms the existing methods on the Stanford2D3D [6] monocular depth estimation benchmark dataset qualitatively and quantitatively.

METHODOLOGY

Our pipeline, shown in Figure 2. consist of two branches. The first coarse level branch takes downsampled, low resolution ERP image as input. While, second fine branch takes high resolution ERP image. These ERP images are projected to multiple perspective tangent images using gnomonic projection. These patches along with their embedded 3D geometric positions are given to encoder network to extract image features. The 3D geometric position embedding is obtained using Multi-Layer Perceptron (MLP) network with tangent image pixel spherical coordinates (λ, ϕ, ρ) on unit sphere and patch center coordinates (λ', ϕ') as input. The bottleneck features are then given to Multi-head Self Attention (MSA) network to obtain long range relationship among tangent image features.

These coarse and fine branch features are then given to Multi-Scale Feature Fusion (MSFF) module that fuses multi-scale tangent features to get updated holistic context guided fine level features. This fused fine level features are then given to fine branch decoder to produce perspective depth maps which are merged to get final ERP depth map. Similar to fine branch, we get coarse depth map by giving coarse level bottleneck feature to decoder network. Also, we refine both depth maps using an iterative approach [5] where the estimated depth values from previous iteration are used to update 3D geometric positions for MLP networks of next iteration. We train our network in an end-to-end manner and use BerHu loss [1] to supervise the estimated depth maps.

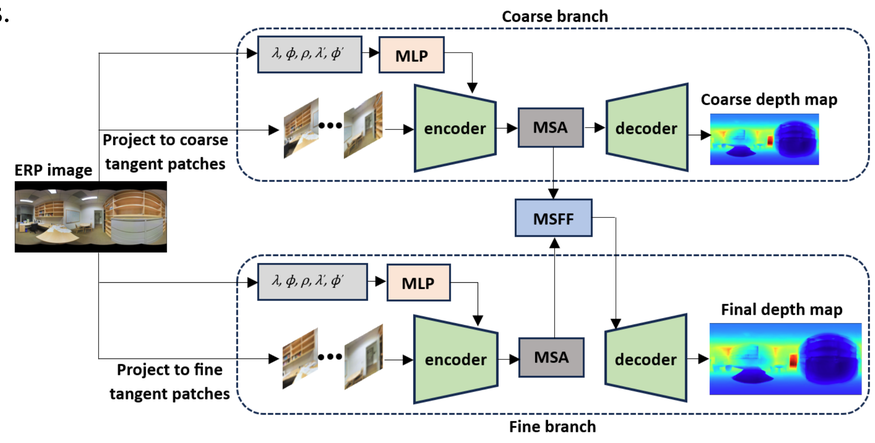

Figure 2. Pipeline of Multi-scale Monocular Panorama Depth Estimation Network

RESULTS

We evaluated the performance of our model using Stanford2D3D [6] monocular depth estimation benchmark dataset. This dataset consists of around 1,413 panorama images of six real world large-scale indoor scenes. We use scene number five for testing, and rest for training. The input resolution of the ERP images for coarse level is 256 x 512 and fine level is 512 x 1024, respectively. For quantitative analysis we used five commonly used metrics called Absolute Relative Error (Abs Rel), Root Mean Squared Error (RMSE) and accuracy with a threshold δ_t where t $\in$ 1.25, 1.25^2, 1.25^3. In this section we show that our model performs best compared to the previous models quantitatively as shown in Table 1. and qualitatively as shown in Figure 3.

Method	Abs Rel↓	RMSE↓	δ_1 ↑	δ_2↑	δ_3 ↑
FCRN [1]	0.1837	0.5774	0.7230	0.9207	0.9731
BiFuse [2]	0.1209	0.4142	0.8660	0.9580	0.9860
UniFuse [3]	0.1114	0.3691	0.8711	0.9664	0.9882
HoHoNet [4]	0.1014	0.3834	0.9054	0.9693	0.9886
OmniFusion [5]	0.0943	0.3582	0.8999	0.9742	0.9914
Our Model	**0.0895**	**0.3423**	**0.9112**	**0.9759**	**0.9921**

Table 1. Quantitative Results on Stanford2D3D [6] Benchmark Dataset

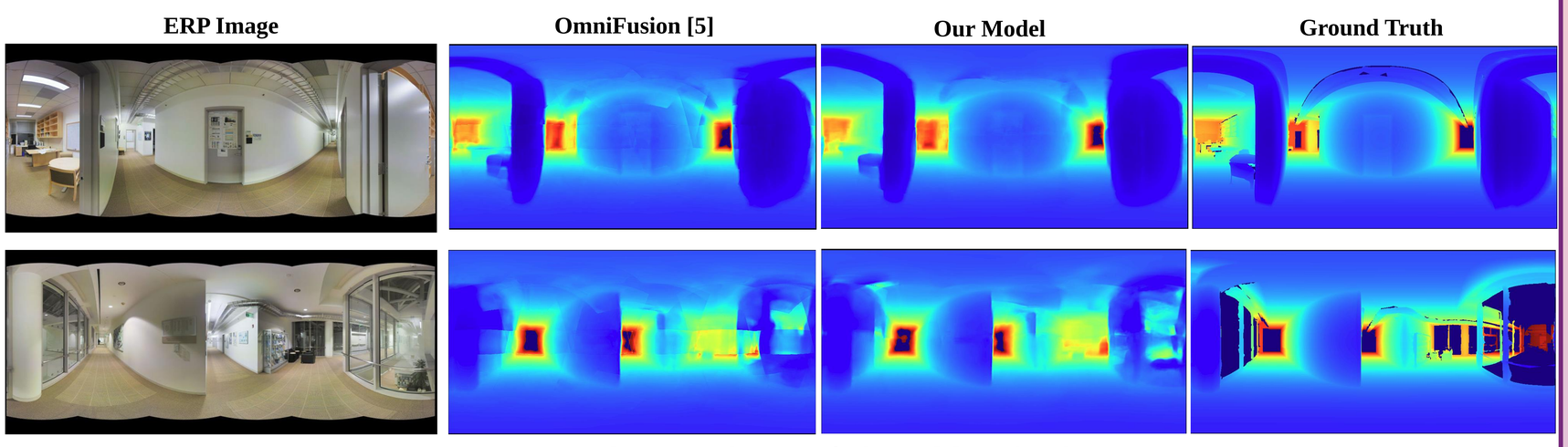

Figure 3. Qualitative Results on Stanford2D3D [6] Benchmark Dataset

Electrical Engineering and Computer Science, University of Missouri

REFERENCES

[1] Iro Laina, C. Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. 2016 Fourth International Conference on 3D Vision (3DV), pages 239–248, 2016.
[2] Fu-En Wang, Yu-Hsuan Yeh, Min Sun, Wei-Chen Chiu, and Yi-Hsuan Tsai. Bifuse: Monocular 360 depth estimation via bi-projection fusion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 462–471, 2020.
[3] Hualie Jiang, Zhe Sheng, Siyu Zhu, Zilong Dong, and Rui Huang. Unifuse: Unidirectional fusion for 360◦ panorama depth estimation. IEEE Robotics and Automation Letters, 2021.
[4] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Hohonet: 360 indoor holistic understanding with latent horizontal features. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 2573–2582, 2021.
[5] Yuyan Li, Yuliang Guo, Zhixin Yan, Xinyu Huang, Ye Duan, and Liu Ren. Omnifusion: 360 monocular depth estimation via geometry-aware fusion. CoRR, abs/2203.00838, 2022.
[6] Iro Armeni, Sasha Sax, Amir R Zamir, and Silvio Savarese. Joint 2d-3d-semantic data for indoor scene understanding. arXiv preprint arXiv:1702.01105, 2017.