360D: A dataset and baseline for dense depth estimation from 360^o images

Antonis Karakottas, Nikolaos Zioulis, Dimitrios Zarpalas, Petros Daras

Centre for Research and Technology Hellas (CERTH) - Information Technologies Institute (ITI) - Visual Computing Lab (VCL)

Abstract. We present a baseline for 360^o dense depth estimation from a single spherical panorama. We circumvent the unavailability of coupled 360^o color and depth image datasets by rendering a high quality 360^o dataset from existing 3D datasets. We then train a CNN designed specifically for 360^o content in a supervised manner, in order to predict a 360^o depth map from a single omnidirectional image in equirectangular format. Quantitative and qualitative results show the need for training directly in 360^o instead of relying on traditional 2D CNNs.

Keywords: Depth Estimation, 360^o, Spherical Panorama

1 Introduction

Omnidirectional (360^o) content is seeing a sudden rise these last few years as new hardware and the progressive maturity of stitching technology allows for easier productions, even extending them to the wider consumer public. Albeit still growing, it remains a new, fresh medium, that allows for interactive and immersive experiences which has greatly benefited from the recent Virtual Reality (VR) advances. Nonetheless, given that there are some fundamental differences

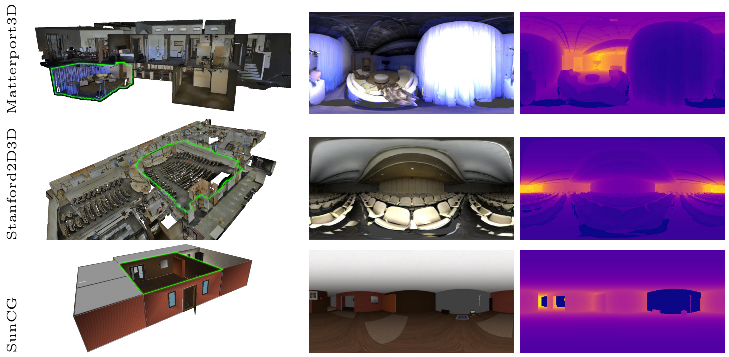

Fig. 1: Sample generated data, from left to right: the 3D scene with a green highlight denoting the rendered room, color output, corresponding depth map.

between 360^o and traditional 2D media, there has been limited activity in transferring the advances in various scene understanding tasks from 2D to 360^o. One such case is monocular dense depth estimation, a very important task for a variety of applications. While recent work has shown impressive results [1, 2], this topic has not been explored at all for the 360^o domain. The main reason for this, is the unavailability of 360^o datasets, be it either in a supervised or unsupervised context with ground truth depth or stereo pairs respectively. For the former case, the coupling of 360^o cameras with 360^o (typically LiDaR) depth sensors is an unfortunate combination due to the differences in resolution and the difficulty in acquiring depth in a full spherical manner. For the latter case, which is also an issue for the former, the second viewpoint (i.e. camera) would be visible from the first one, and vise versa, making most reasonable baselines problematic in terms of acquiring high quality data. In this work, we circumvent the lack of proper datasets, by generating a semi-synthetic one, harnessing the availability of large scale 3D indoor scene datasets, both synthetic(i.e. computer generated) or realistic (i.e. scanned buildings). Additionally we design a CNN for dense depth prediction trained on this dataset to serve as a baseline for future research. We offer both the 360^o data and trained models in http://vcl.iti.gr/360-dataset/.

2 360D Dataset

We generate a dataset of 360^o indoor scenes with their corresponding depth annotations by rendering indoor scenes from recent large scale datasets of textured 3D scenes. We rely on realistic scenes acquired by scanning actual buildings as well as synthetic computer generated ones. For the former we use the Stanford2D3D [3] and Matterport3D [4] datasets, while for the latter we use the SunCG [5] dataset. In order to render in 360^o, we utilize the ray tracing engine available in Blender. We use a uniform point light source positioned in the same position as the camera. Each render offers a rendered color image and its corresponding ground truth depth map extracted from the z-buffer. Our generated 360D dataset contains a total of 23524 unique viewpoints, of synthetic and real-

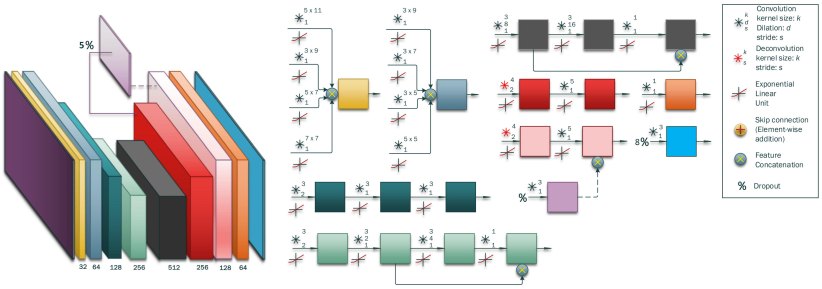
5%
32
64
128
256
512
256
128
64
5 x 11
3 x 9
5 x 7
7 x 7
3 x 9
3 x 7
3 x 5
5 x 5
8%
%
Convolution kernel size: k Dilation: d stride: s
Deconvolution kernel size: k stride: s
Exponential Linear Unit
Skip connection (Element-wise addition)
Feature Concatenation
% Dropout

Fig. 2: Our proposed 360^o CNN architecture decomposed in blocks.

istic 360^o color and depth image data in a variety of indoors contexts (houses, offices, etc.) in various layouts, with samples shown in Fig. 1.

3 Omnidirectional Depth Estimation

We design a CNN architecture specifically for 360^o learning taking into account two important differences compared to traditional images. Omnidirectional images, when in equirectangular format, suffer from high distortions along their vertical axis which increases towards the spheres poles and is also different for every image row. Therefore, information is scattered horizontally, as we vertically approach the two poles. We utilize rectangle filters of varying width following the work of [6] where the CNN's 2D filters are transferred into distorted row-wise versions that approximate rectangle shapes. This is shown to increase performance when applied to the 360^o domain. A block consisting of the concatenation of a traditional square filter and three rectangle ones are used for the first two convolutional layers, with the rectangle filter sizes chosen so as to preserve the area of the square filter.

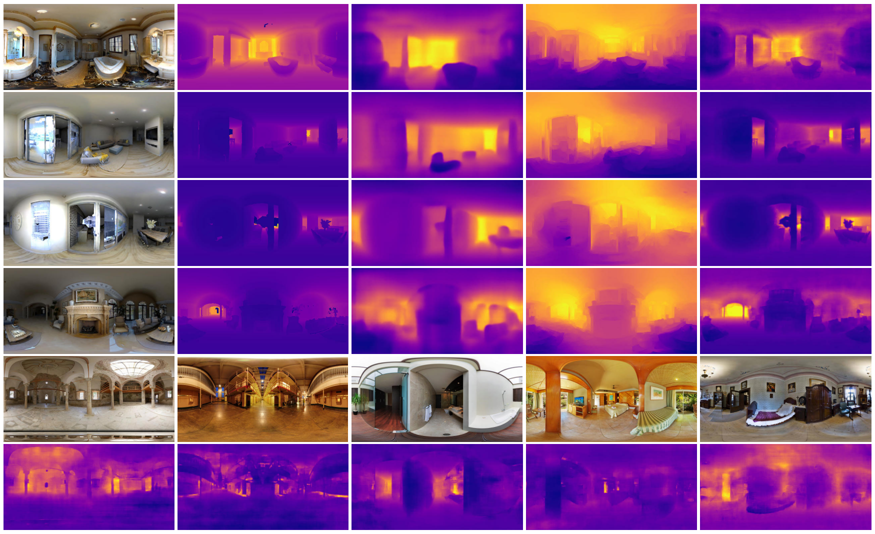

Fig. 3: Qualitative results on our test split and on samples of the Sun360 dataset [7]. First four rows from left to right: color image, ground truth depth and predictions of Laina et al. [1], Liu et al. [2], ours. Last two rows from top to bottom: unseen realistic samples from the Sun360 [7] dataset, our depth predictions.

Further, unlike 2D images, 360^o content captures a scene's global context. We exploit this opportunity offered by spherical content and design our network architecture with the goal of maximizing its receptive field (RF) utilizing dilated convolutions [8]. At the same time, we preserve the input images' spatial resolution as much as possible by using progressively increasing dilations instead of progressive downscaling which diminishes the scene's structural details. We derive this inspiration from [9] where this technique was shown to perform better in a global scene understanding task. Fig. 2 presents our architecture in detail.

We evaluate our 360^o depth estimation network's performance by offering quantitative comparisons against existing monocular depth estimation methods - given the unavailability of 360^o networks for this task - in Table 1, using standard metrics as those found in [1, 2]. In addition we present qualitative results in unseen, realistic data in Fig. 3.

Table 1: Comparison of our CNN performance to Laina et al. [1] and Liu et al. [2]. (arrows denote direction of better performance)

Network	Abs Rel ↓	Sq Rel ↓	RMS ↓	RMS(log) ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
Ours	**0.0702**	**0.0297**	**0.2911**	**0.1017**	**0.9574**	**0.9933**	**0.9979**
Laina et al. [1]	0.3181	0.4469	0.941	0.376	0.4922	0.7792	0.915
Liu et al. [2]	0.4202	0.7597	1.1596	0.44	0.3889	0.7044	0.8774

References

1. Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., Navab, N.: Deeper depth prediction with fully convolutional residual networks. In: 3D Vision (3DV), 2016 Fourth International Conference on, IEEE (2016) 239–248
2. Liu, F., Shen, C., Lin, G., Reid, I.: Learning depth from single monocular images using deep convolutional neural fields. IEEE transactions on pattern analysis and machine intelligence **38**(10) (2016) 2024–2039
3. Armeni, I., Sener, O., Zamir, A.R., Jiang, H., Brilakis, I., Fischer, M., Savarese, S.: 3d semantic parsing of large-scale indoor spaces. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2016) 1534–1543
4. Chang, A., Dai, A., Funkhouser, T., Halber, M., Niessner, M., Savva, M., Song, S., Zeng, A., Zhang, Y.: Matterport3D: Learning from RGB-D data in indoor environments. International Conference on 3D Vision (3DV) (2017)
5. Song, S., Yu, F., Zeng, A., Chang, A.X., Savva, M., Funkhouser, T.: Semantic scene completion from a single depth image. IEEE Conference on Computer Vision and Pattern Recognition (2017)
6. Su, Y.C., Grauman, K.: Learning spherical convolution for fast features from 360 imagery. In: Advances in Neural Information Processing Systems. (2017) 529–539
7. Xiao, J., Ehinger, K.A., Oliva, A., Torralba, A.: Recognizing scene viewpoint using panoramic place representation. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE (2012) 2695–2702
8. Yu, F., Koltun, V., Funkhouser, T.: Dilated residual networks. In: Computer Vision and Pattern Recognition. Volume 1. (2017)
9. van Noord, N., Postma, E.O.: Light-weight pixel context encoders for image inpainting. CoRR **abs/1801.05585** (2018)