SliceNet: deep dense depth estimation from a single indoor panorama using a slice-based representation: Supplementary Material

Giovanni Pintore
Visual Computing, CRS4, Italy
giovanni.pintore@crs4.it

Marco Agus
CSE, HBKU, Doha, Qatar
magus@hbku.edu.qa

Eva Almansa
Visual Computing, CRS4, Italy
evaalmansa@crs4.it

Jens Schneider
CSE, HBKU, Doha, Qatar
jeschneider@hbku.edu.qa

Enrico Gobbetti
Visual Computing, CRS4, Italy
enrico.gobbetti@crs4.it

1. Introduction

This supplementary material accompanies the presentation of our method with additional information not included in the main article [5].

First of all, we provide a detailed illustration of the structure of our network architecture (Sec. 2). This illustrations provides details on all the individual network components and is aimed to complement the general description provided in the paper.

Second, we provide a detailed gravity-alignment study (Sec 3) that shows that available benchmark datasets are all well-aligned with respect to the gravity vector and that our method is robust to small gravity misalignments. These additional results show that our method can be directly applied in practice, even without recurring to pre-processing [1].

2. Detailed network architecture description

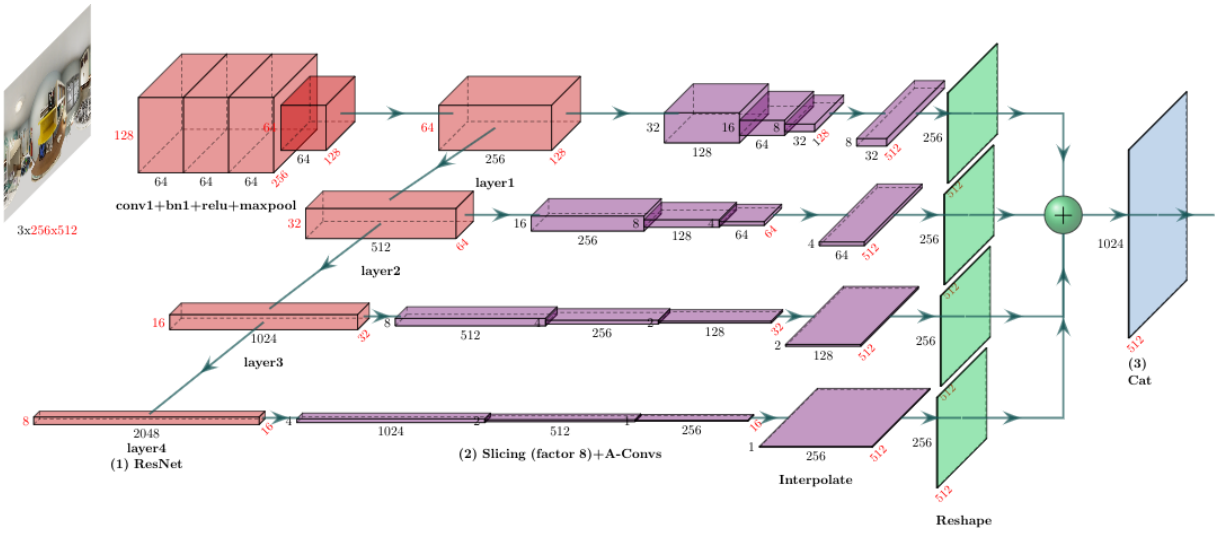
Our deep convolutional neural network (CNN) architecture takes as input an equirectangular RGB image and outputs a registered depth image at the same resolution of the input. The detailed structure of the network is illustrated in Fig. 1. The network uses an encoder/decoder structure. The encoder is presented in Fig. 1(a), while the decoder is presented in Fig. 1(b).

The first 8 layers of the network consist of a standard ResNet encoder (Fig. 1(a)). The results presented in the paper are obtained with a ResNet50, but we verified that very good performances can also be obtained and with ResNet18 and ResNet34, with a considerable increase in terms of speed. The last 4 levels of the encoder are sliced, keeping the horizontal dimension unchanged and compressing the vertical one. This way, we accumulate a series of features associated with each element of the horizontal dimension (i.e., a slice). In order to merge the features, coming from different resolution levels and associated to the same slice, we interpolate the 4 maps so that they have the same horizontal dimension (i.e., 512). We then reshape and concatenate the 4 maps so as to obtain a single-sequential bottleneck (i.e., 1024×512).

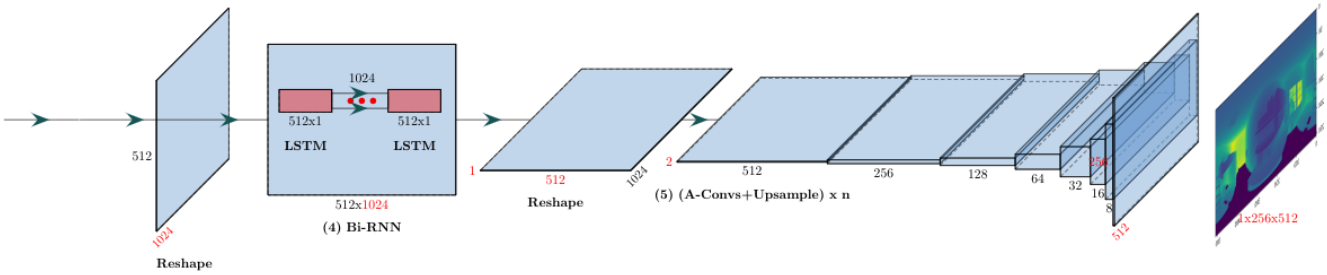
The decoder (Fig. 1(b)) exploits a bi-directional LSTM with 512 hidden layers, which outputs a time-step of size 2×512 for each of the 512 slices. So, that the final output of this block is a feature map having the same size of the RNN block input, i.e., 1024×512. Once reshaped to $1024 \times 1 \times 512$, this flattened representation is upsampled to the desired output size (i.e., $1 \times 256 \times 512$) by following steps symmetrical to those used for encoding reduction.

3. Detailed gravity-alignment study

Our approach starts from the assumption that gravity plays an important role in the design and construction of interior environments, and that world-space vertical and horizontal features have different characteristics in most, if not all, man-made environments. Based on this fact, we strive to exploit gravity-aligned world-space features by performing a gravity-aligned processing of images. This assumes that input equirectangular images are aligned to the gravity vector. While this assumption could be managed by gravity-aligning images before our pipeline, it is rational to assume that, in most cases, captured images already meet these constraints. To verify this fact, we performed a study of gravity-alignment of available datasets, and verified the robustness of our method to small misalignment.



(a) *Encoder*



(b) *Decoder*

Figure 1. **Detailed illustration of the SliceNet architecture.** This illustration complements the architectural view provided in the paper. The network uses an encoder/decoder structure. The encoder is presented in Fig. 1(a), while the decoder is presented in Fig. 1(b). The last 4 levels of the encoder are sliced, keeping the horizontal dimension unchanged and compressing the vertical one (Fig. 1(a)). From the resulting sliced sequence ($1024 \times 1 \times 512$), we recover long and short term information through a LSTM module (Fig. 1(b)). The final depth map is recovered by following steps symmetrical to those used for encoding reduction.

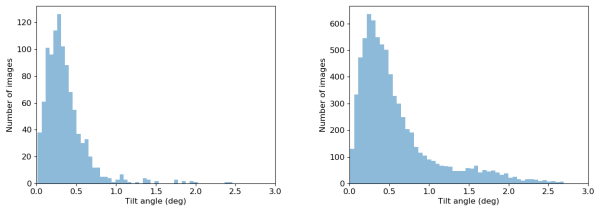
3.1. Gravity-alignment evaluation of benchmark datasets

All the commonly available synthetic datasets [9, 8] are perfectly aligned by design, and they thus perfectly meet the constraint.

The study, thus, focuses on real-world capture. A common practice for capturing an indoor scene is to place the camera on a tripod placed on a horizontal plane [1]. This capture method is in fact adopted in all the datasets available for benchmarking and also adopted in our work and the compared state-of-the-art methods [2, 10, 7].

For real-world datasets [6, 4] we exploited the alignment pipeline of Zou et al. [11] to evaluate the misalignment with the ground plane (see Fig. 2).

In our experiments we found that the average inclination, with respect to the gravity vector, is 0.36 degrees for the Stanford2D3D [6] dataset, while the average misalignment of the Matterport3D [4] dataset is 0.61 degrees.



(a) *Stanford2D3D* (b) *Matterport3D*

Figure 2. **Real-world datasets vertical misalignment.** The average inclination with respect to the gravity vector of the Stanford2D3D [6] dataset is about 0.36 degrees, while the average misalignment of the Matterport3D [4] dataset is about 0.61 degrees. Outliers are mainly due to inaccurate line detection and classification of the alignment tool [3].

Indeed these values are really minimal, also considering that a significant part of the angular error is due to low ac-

curacy detecting lines and estimating dominant direction by the automatic alignment tool [3]. We can, therefore conclude that available datasets all have a sub-degree accuracy with respect to gravity alignment.

3.2. Robustness to gravity misalignment

Even if our method assumes to work with gravity-aligned scenes, we do not necessary require a perfect alignment. In addition to the results and comparison already presented in the paper, we show, for completeness, the results obtained by introducing various degrees of error in the alignment ($0°$, $\pm 2°$, $\pm 5°$). We also performed a test, combining both *training* and *testing* of Structured3D [8] with and without alignment to the ground plane.

Results in Tab. 1 demonstrate the consistency of our model and effectiveness of our assumption, where the best performances are obtained the more the images are aligned with the ground plane, while the results do not improve even if a specific training is done on distorted data in order to find a better fit on the inclined images. Moreover, the method appears fairly robust to small alignment errors ($\leq \pm 2°$), and degrades as soon as input images are severely misaligned.

Train incl.	Test incl.	MRE	MAE	RMSE	RMSE log	δ_1
$0°$	$0°$	0.0147	0.1180	0.0549	0.1012	0.9085
$0°$	$\pm 2°$	0.0217	0.1393	0.0658	0.1368	0.8776
$0°$	$\pm 5°$	0.0263	0.1601	0.0714	0.1430	0.8527
$\pm 2°$	$0°$	0.0238	0.1516	0.0632	0.1288	0.8672
$\pm 2°$	$\pm 2°$	0.0250	0.1589	0.0716	0.1434	0.8464
$\pm 2°$	$\pm 5°$	0.0281	0.1716	0.0743	0.1501	0.8310
$\pm 5°$	$0°$	0.0231	0.1530	0.0648	0.1245	0.8638
$\pm 5°$	$\pm 2°$	0.0250	0.1613	0.0721	0.1388	0.8438
$\pm 5°$	$\pm 5°$	0.02758	0.1697	0.0735	0.01422	0.8362

Table 1. **Performance when training with misaligned images.** We show, for completeness, the results obtained by combining both training and testing with and without alignment to the ground plane on the Structured3D dataset [8].

In other words, the effectiveness of the network is not given by the specific fitting of the training data with the expected result but by the consistency of the scene with our specific network architecture.

Acknowledgments This work received funding from the European Union's H2020 research and innovation programme under grant 813170 (EVOCATION), and from Sardinian Regional Authorities under project VIGECLAB (POR FESR 2014-2020).

References

[1] Benjamin Davidson, Mohsan S. Alvi, and Joao F. Henriques Henriques. 360 camera alignment via segmentation. In *Proc. ECCV*, pages 579–595, 2020.

[2] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab. Deeper depth prediction with fully convolutional residual networks. In *Proc. 3DV*, pages 239–248, 2016.

[3] David C Lee, Martial Hebert, and Takeo Kanade. Geometric reasoning for single image structure recovery. In *Proc. CVPR*, pages 2136–2143, 2009.

[4] Matterport. Matterport3D. https://github.com/niessner/Matterport, 2017. [Accessed: 2019-09-25].

[5] Giovanni Pintore, Marco Agus, Eva Almansa, Jens Schneider, and Enrico Gobbetti. SliceNet: deep dense depth estimation from a single indoor panorama using a slice-based representation. In *Proc. CVPR*, 2021. To appear.

[6] Stanford University. BuildingParser Dataset. http://buildingparser.stanford.edu/dataset.html, 2017. [Accessed: 2019-09-25].

[7] 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 *Proc. CVPR*, June 2020.

[8] Jia Zheng, Junfei Zhang, Jing Li, Rui Tang, Shenghua Gao, and Zihan Zhou. Structured3D: A large photo-realistic dataset for structured 3D modeling. In *Proc. ECCV*, pages 519–535, 2020.

[9] Nikolaos Zioulis, Antonis Karakottas, Dimitris Zarpalas, Federic Alvarez, and Petros Daras. Spherical view synthesis for self-supervised 360° depth estimation. In *Proc. 3DV*, September 2019.

[10] Nikolaos Zioulis, Antonis Karakottas, Dimitrios Zarpalas, and Petros Daras. OmniDepth: Dense depth estimation for indoors spherical panoramas. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Proc. ECCV*, pages 453–471, 2018.

[11] Chuhang Zou, Alex Colburn, Qi Shan, and Derek Hoiem. LayoutNet: Reconstructing the 3D room layout from a single RGB image. In *Proc. CVPR*, pages 2051–2059, 2018.