MultiPanoWise: holistic deep architecture for multi-task dense prediction from a single panoramic image

Uzair Shah, Muhammad Tukur, Mahmood Alzubaidi
ICT Division, College of Science and Engineering, Hamad Bin Khalifa University
Doha (Qatar)

Giovanni Pintore, Enrico Gobbetti
Visual and Data-intensive Computing, CRS4, Italy
National Research Center in HPC, Big Data, and Quantum Computing, Italy
(giovanni.pintore|enrico.gobbetti)@crs4.it

Mowafa Househ, Jens Schneider, Marco Agus
ICT Division, College of Science and Engineering, Hamad Bin Khalifa University
Doha (Qatar)
(magus|jeschneider)@hbku.edu.qa

Abstract

We present a novel holistic deep-learning approach for multi-task learning from a single indoor panoramic image. Our framework, named MultiPanoWise*, extends vision transformers to jointly infer multiple pixel-wise signals, such as depth, normals, and semantic segmentation, as well as signals from intrinsic decomposition, such as reflectance and shading. Our solution leverages a specific architecture combining a transformer-based encoder-decoder with multiple heads, by introducing, in particular, a novel context adjustment approach, to enforce knowledge distillation between the various signals. Moreover, at training time we introduce a hybrid loss scalarization method based on an augmented Chebychev/hypervolume scheme. We illustrate the capabilities of the proposed architecture on public-domain synthetic and real-world datasets. We demonstrate performance improvements with respect to the most recent methods specifically designed for single tasks, like, for example, individual depth estimation or semantic segmentation. To our knowledge, this is the first architecture capable of achieving state-of-the-art performance on the joint extraction of heterogeneous signals from single indoor omnidirectional images.*

Additional qualitative results

In the following tables we showcase enlarged versions of the manuscript images together with additional qualitative comparisons with PanoFormer [2] on synthetic Structured3d [3] and real world Stanford2d3d [1] datasets.

References

[1] 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. 1

[2] Zhijie Shen, Chunyu Lin, Kang Liao, Lang Nie, Zishuo Zheng, and Yao Zhao. Panoformer: Panorama transformer for indoor 360 depth estimation. In *Computer Vision – ECCV 2022*, pages 195–211, Cham, 2022. Springer Nature. 1, 2, 3, 4

[3] Jia Zheng, Junfei Zhang, Jing Li, Rui Tang, Shenghua Gao, and Zihan Zhou. Structured3d: A large photo-realistic dataset for structured 3d modeling. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 519–535. Springer, 2020. 1

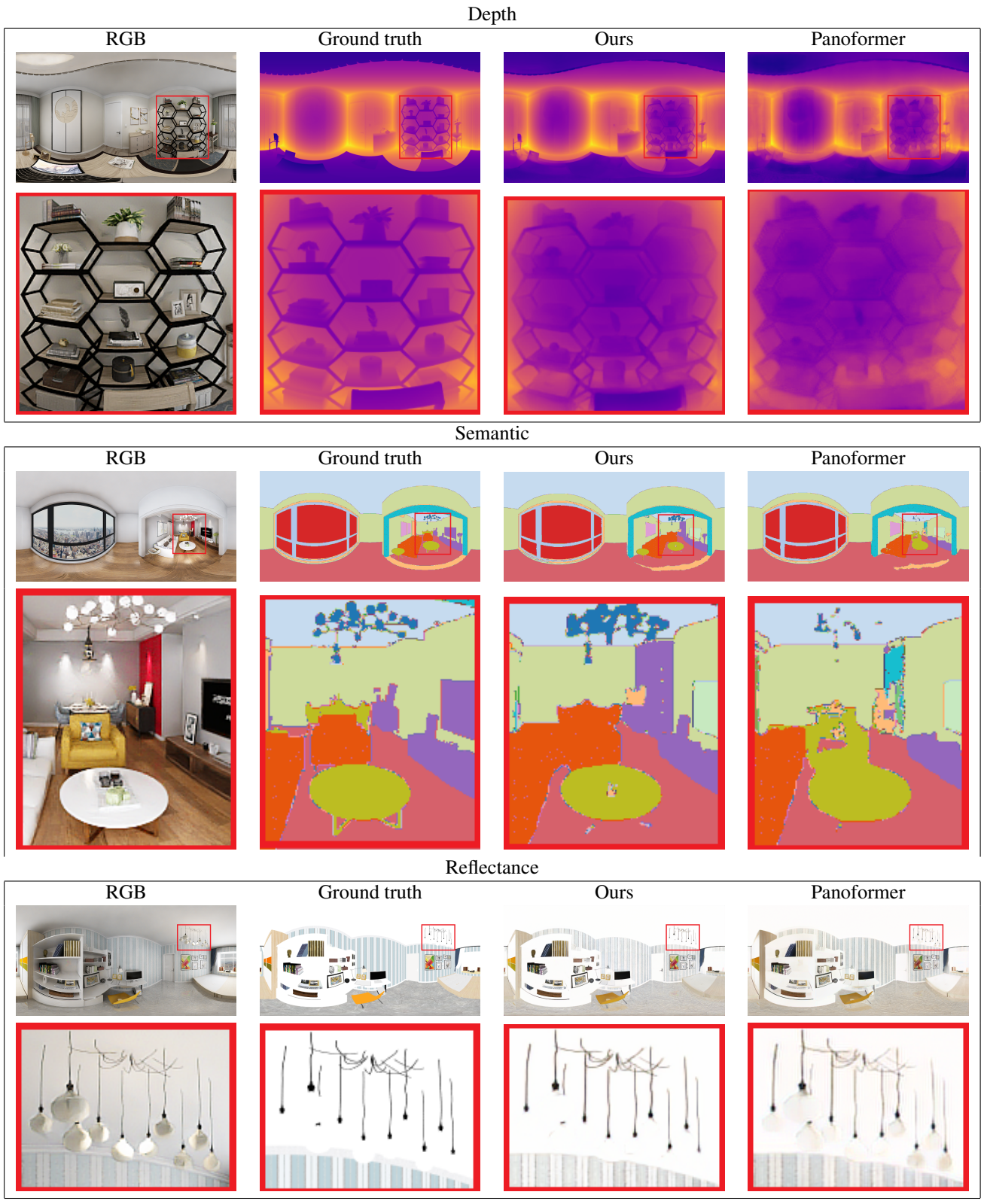

Table 1. Qualitative comparison with PanoFormer [2] for single inference on Structured3d dataset for depth, semantic, and albedo

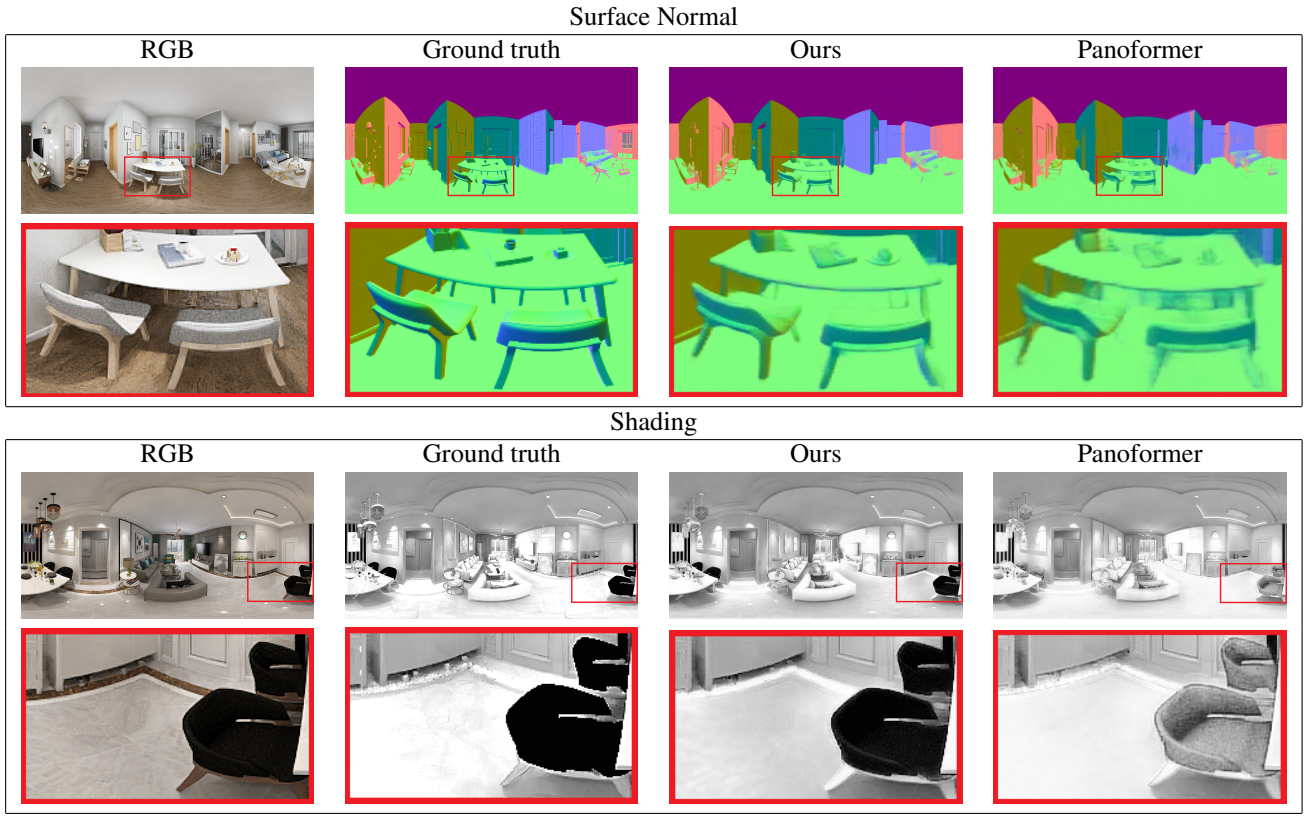

Table 2. Qualitative comparison with PanoFormer [2] for single inference on Structured3d dataset for normal and shading

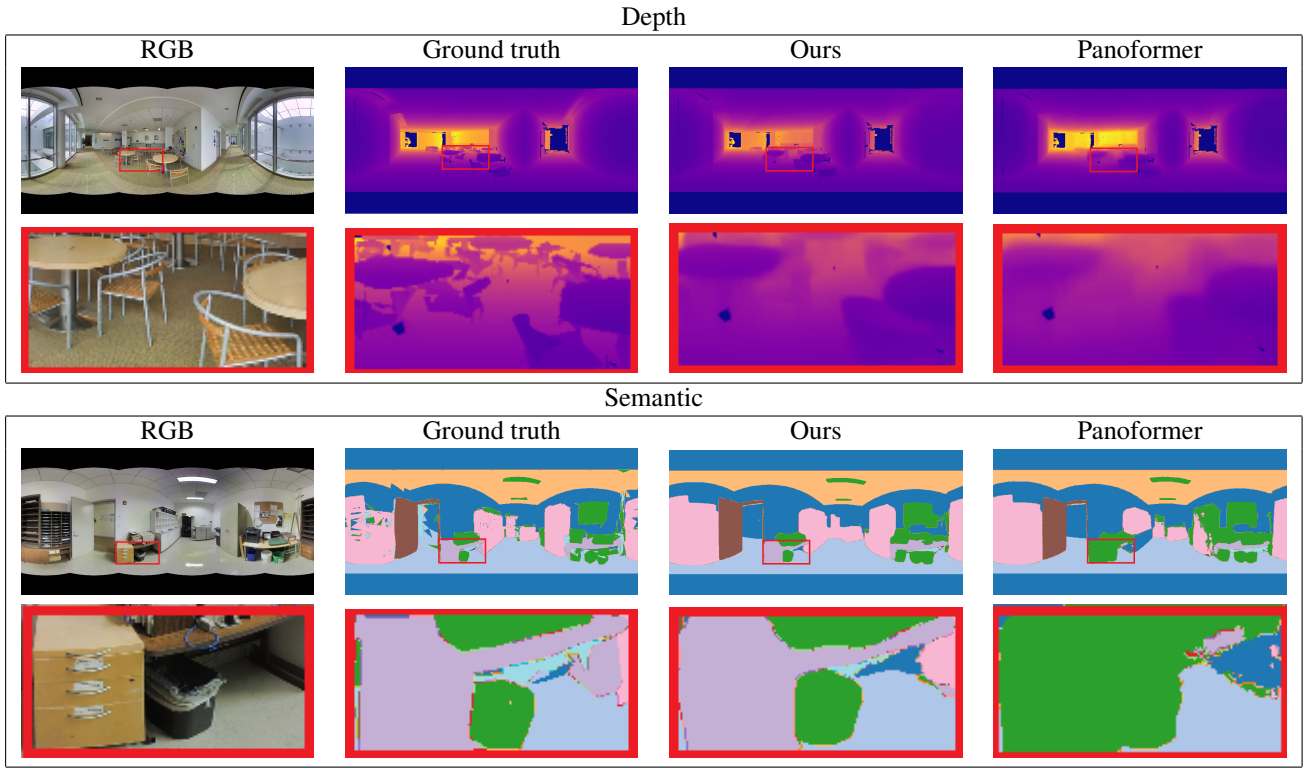

Table 3. Qualitative comparison with PanoFormer [2] for single inference on Structured3d dataset for depth and semantic