Relationship Spatialization for Depth Estimation

Xiaoyu Xu[1], Jiayan Qiu[1], Xinchao Wang[2], and Zhou Wang[1]

[1] University of Waterloo, Canada
[2] National University of Singapore, Singapore
x423xu@uwaterloo.ca, jiayan.qiu.1991@outlook.com, xinchao@nus.edu.sg, zhou.wang@uwaterloo.ca

1 Supplementary materials

Due to the page limitation of the main paper, we show visualization effects of the proposed method, and state-of-the-art methods Adabins [1], Midas [6], Bts [5] in supplementary materials. Besides, we present that if only the relationship representations are used, the proposed model also estimate a coarse depth map, which further proves the relationship representations embed with spatial priors.

1.1 Comparison

In this section, we present the visualized results of the enhanced baseline model and state-of-the-art models against original baseline model and state-of-the-art models. The test datasets include KITTI [2], NYU Depth v2 [7] and ICL_NUIM [3].

1.2 Tested on visual-genome [4]

In this section, we test the model performance on visual-genome dataset [4]. In this dataset, the relationship recognition model performs better, so the depth estimation is more accurate. The used model is enhanced Adabins [1].

1.3 User Study

User study for respective contributions quantification Due to the lack of ground-truth annotations of the respective contributions, a user study is designed and performed to evaluate the quantified respective contributions from our framework. Specifically, we involve 107 users and involve each user in 100 multi-round tests. As shown in Figure. 5, for the first round of each multi-round test we first locate a target object in the scene, table in this case, and then erase one of its identified effective relationships Figure. 5(b-d), finally, we show the original scene image and its erased ones to the user. At the end of this round, the user is asked to choose the most important relationship for estimating the target object. Then in the next round, the erased image from the previous round that erases the relationship with the highest respective contribution is used as the ground-truth image. Afterward, we further erase one of the remaining effective

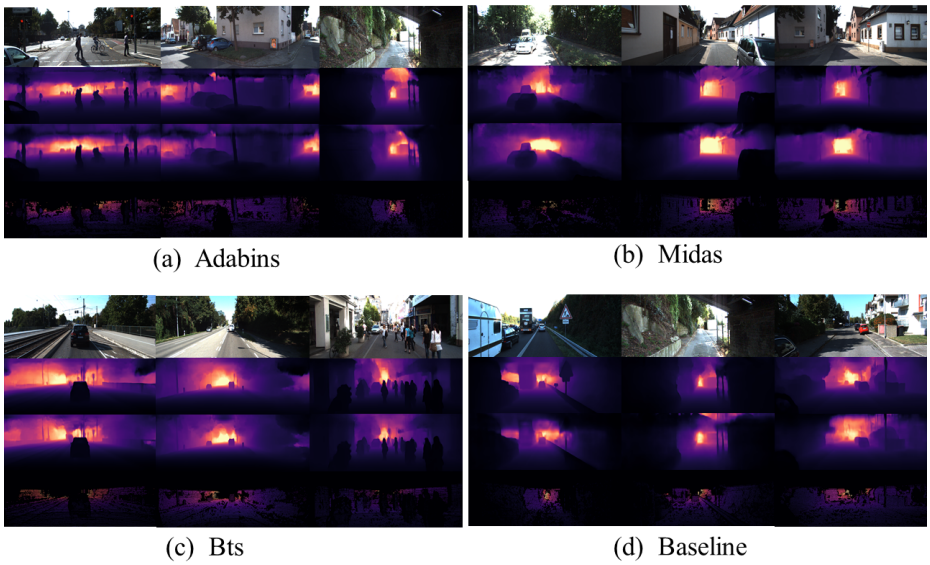

Fig. 1: Visual examples on KITTI [2] test data. The first row: input images; The second row: relationships-enhanced depth maps; The third row: depth maps without relationships information; The last row: ground-truth depth maps.

relationships, Figure. 5(e-g), and ask the user to choose the highest contributed relationship in this situation. Finally, we evaluate the accuracy of our respective contributions according to the rank of the users. As can be seen from Table 1, for target objects, our top-5 accuracy is around 91.25%, which shows that our learned respective contribution is highly aligned with the human recognition.

Metrics	Top 1 Acc	Top 2 Acc	Top 3 Acc	Top 4 Acc	Top 5 Acc
(%)	86.78	87.23	89.20	90.44	91.25

Table 1: Quantification of relationship effectiveness.

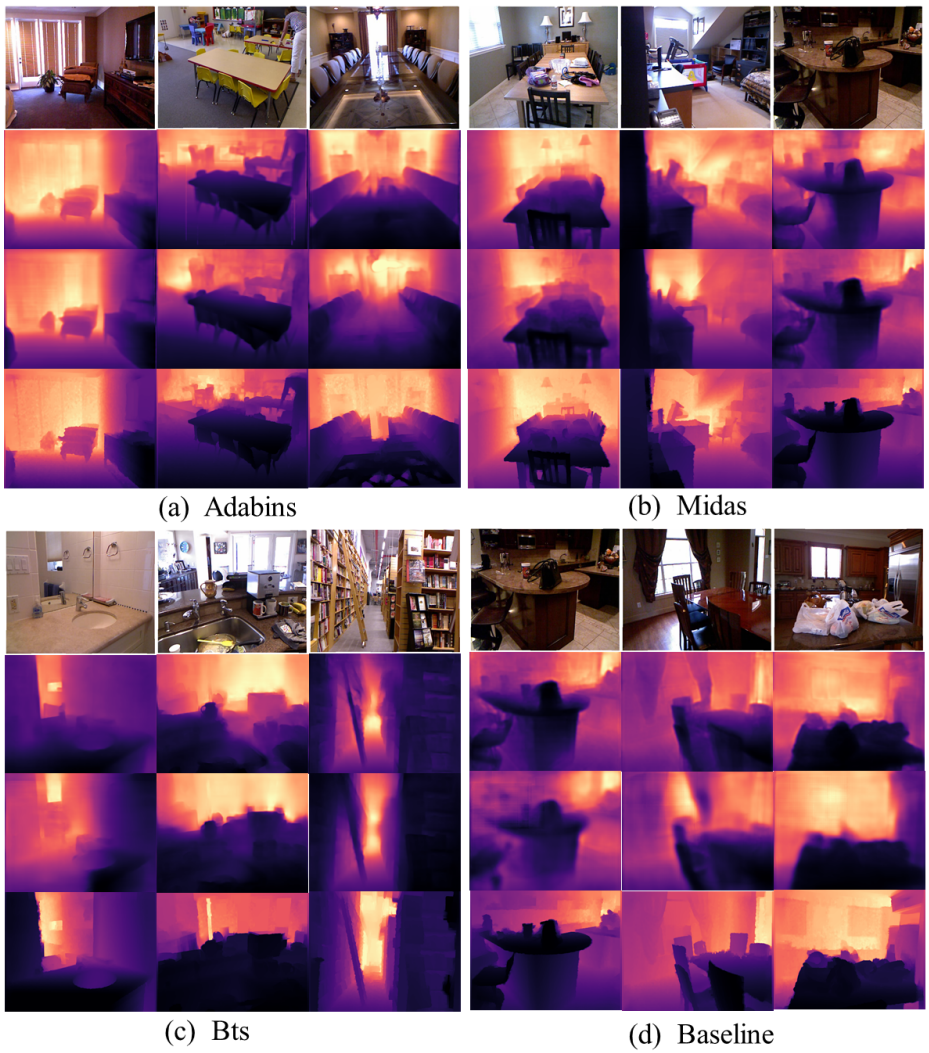

Fig. 2: Visual examples on NYU Depth v2 [7] test data. The first row: input images; The second row: relationships-enhanced depth maps; The third row: depth maps without relationships information; The last row: ground-truth depth maps.

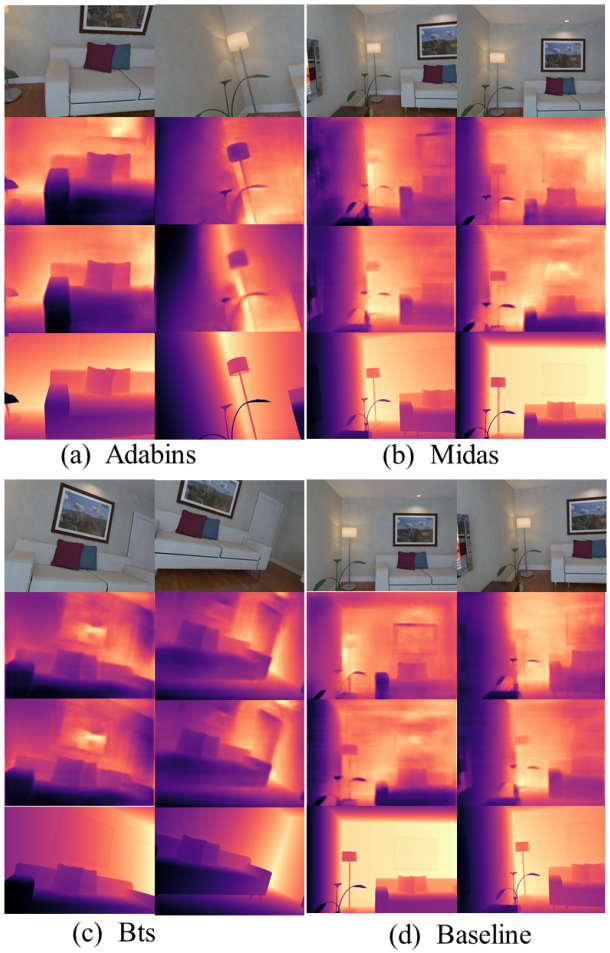

Fig. 3: Visual examples on ICL_NUIM [3] test data. The first row: input images; The second row: relationships-enhanced depth maps; The third row: depth maps without relationships information; The last row: ground-truth depth maps.

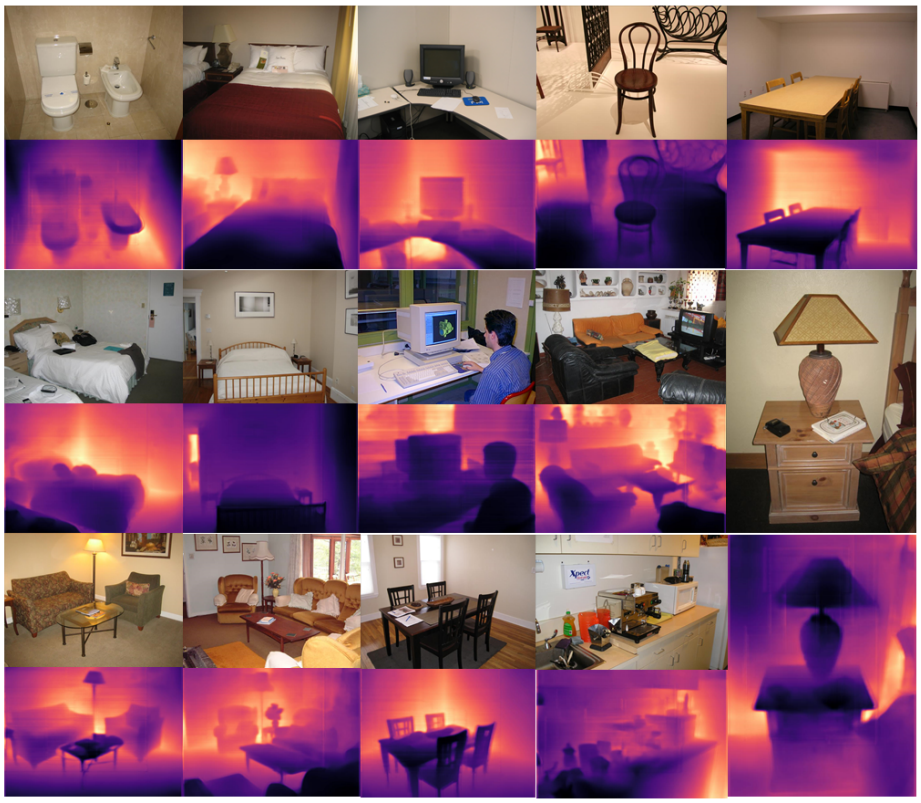

Fig. 4: Visual examples on visual genome [4].

Fig. 5: Quantification of relationships effectiveness. (a): ground truth image; (b)-(d): In the first round, erase each object and select the most important relationship; (e)-(g): In the second round, use image with the selected object erased as ground truth and select the next most important object.

References

1. Bhat, S.F., Alhashim, I., Wonka, P.: Adabins: Depth estimation using adaptive bins. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4009–4018 (2021)
2. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. The International Journal of Robotics Research **32**(11), 1231–1237 (2013)
3. Handa, A., Whelan, T., McDonald, J., Davison, A.: A benchmark for RGB-D visual odometry, 3D reconstruction and SLAM. In: IEEE Intl. Conf. on Robotics and Automation, ICRA. Hong Kong, China (May 2014)
4. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International journal of computer vision **123**(1), 32–73 (2017)
5. Lee, J.H., Han, M.K., Ko, D.W., Suh, I.H.: From big to small: Multi-scale local planar guidance for monocular depth estimation. arXiv preprint arXiv:1907.10326 (2019)
6. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. arXiv preprint arXiv:1907.01341 (2019)
7. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: European conference on computer vision. pp. 746–760. Springer (2012)