Video-Based, Occlusion-Robust Multi-View Stereo Using Inner-Boundary Depths of Textureless Areas

Jian Wei, Shigang Wang, Yan Zhao

College of Communication Engineering, Jilin University, Changchun, China

weijian@jlu.edu.cn, wangshigang@vip.sina.com, zhao_y@jlu.edu.cn

Motivation

Multi-View Stereo (MVS) for unstructured videos:

Advantages over traditional MVS:

- Capturing a video is much simpler and faster than taking multiple photographs.
- Densely sampled input enables accurate recovery of 3D edges (object boundaries and textured regions).

Problems shared with traditional MVS:

- Severer occlusions and weak textures normally lead to unreliable depth estimates.
- High-level techniques are computationally expensive, particularly for high spatio-temporal resolution videos.

Contribution

A video-based MVS approach is introduced, that efficiently addresses both occlusions and lack of textures.

- We exploit reliable edge depths to recover inner boundaries of visible textureless areas, which are used to infer dense geometry without wrong connections between objects.
- Our approach respects two local cues with complementary advantages, *i.e.* smoothness and density of reconstructed 3D surfaces.
- The algorithm only relies on low-level techniques, *e.g.* intra-view depth interpolation and inter-view depth propagation. Most operations are evaluated per pixel, thus supporting parallel execution.

Method

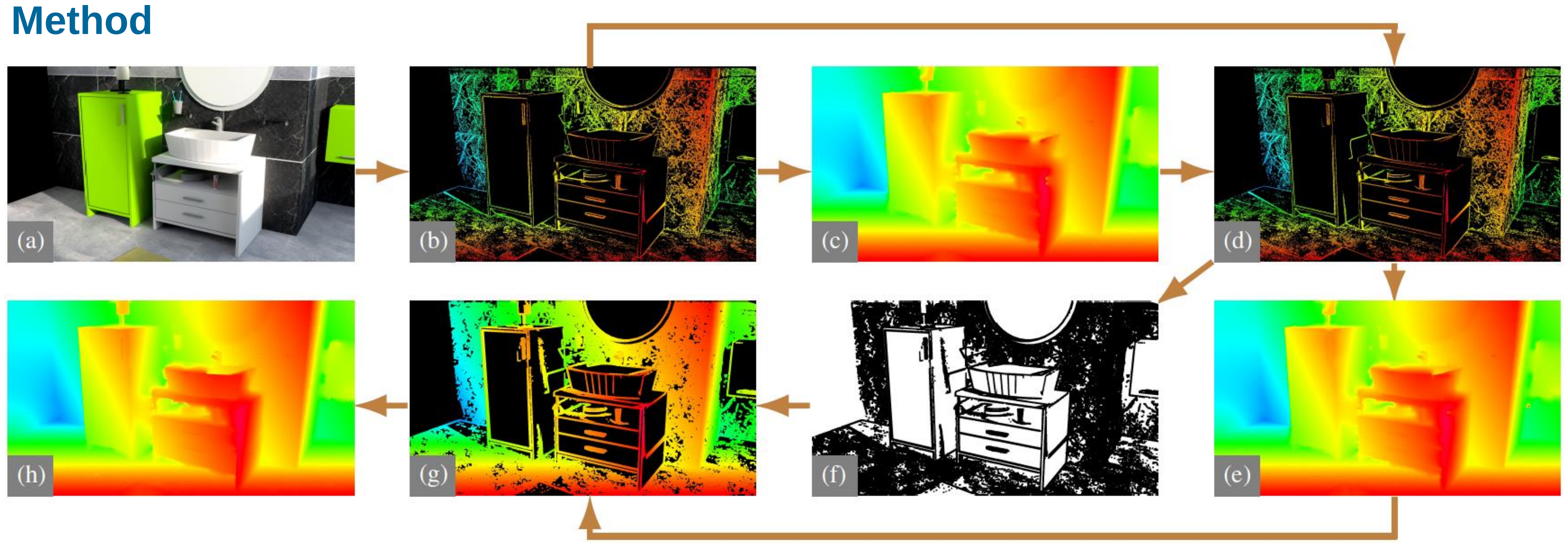

Fig. 1. Flowchart of our approach. Given a video, we first execute SfM to compute camera extrinsics, select a dense frame subset $\{I_i\}$, and calculate edge depths for each I_i (a) using the techniques of [3]. Then, the edge depth maps $\{D_i^e\}$ (b) are refined to improve continuity (c, d), and afterwards leveraged to derive the depth maps $\{D_i^b\}$ for inner boundaries of interior textureless areas (e~g). Dense depth maps (h) are finally obtained through per-view interpolation of $\{D_i^b\}$.

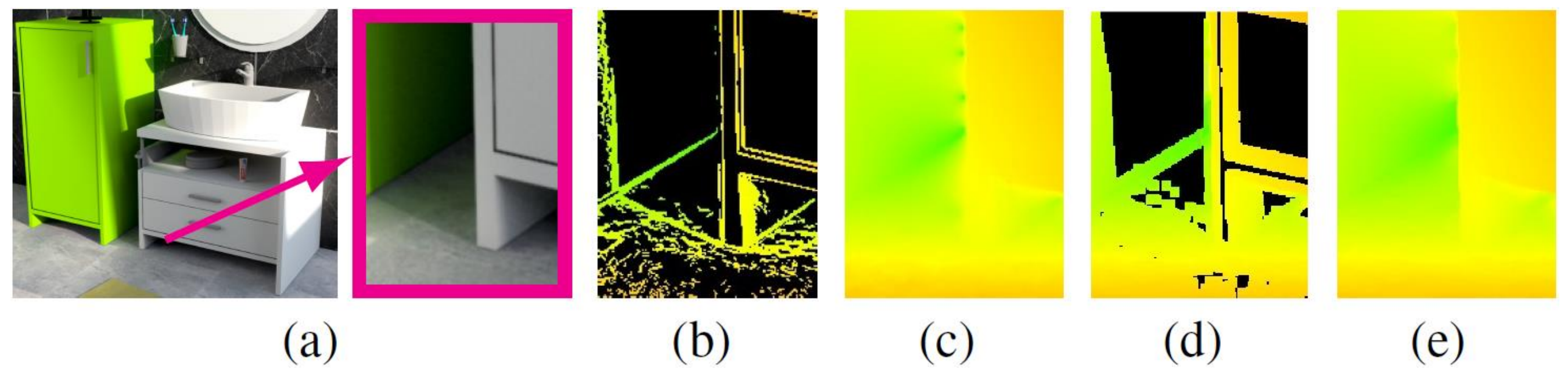

Fig. 2. Depth calculation combining the smoothness and density cues of surfaces. (a) Image area. (b) Edge depths. (c) Interpolated depths from (b). (d) Inner-boundary depths of visible textureless areas. (e) Interpolated depths from (d).

Only considering surface smoothness induces incorrect interpolants, most of which have much lower density than others. But simply selecting the densest surfaces among all views probably produce layered geometry in weakly textured areas. Therefore, we merely depend on surface density at inner boundaries of textureless areas, where the depths obtained in such a way are usually reliable. Then we infer dense geometry under the smooth-surface assumption again.

Results

Our experiments used three videos, all with large textureless areas and arbitrary camera trajectories. Three MVS methods were compared: BAI [1], KIM [2], and WEI [3]. Evaluations were run on multithreaded CPUs and GPUs.

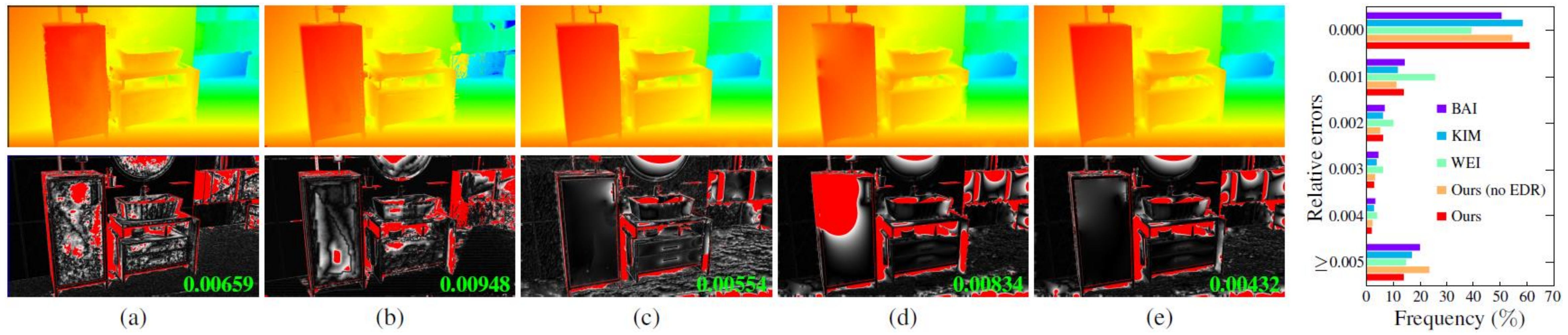

Fig. 3. Depth maps (top), relative error (RE) maps against the ground truth (bottom), and RE distributions (right) obtained by (a) BAI, (b) KIM, (c) WEI, and ours (d) without and (e) with the edge depth refinement step. In the RE maps, the red pixels have a RE larger than 0.01, the REs between 0 and 0.01 are marked gray 0 to 255, and the mean RE of each method is labeled in the lower right corner. Our approach achieves the highest overall precision.

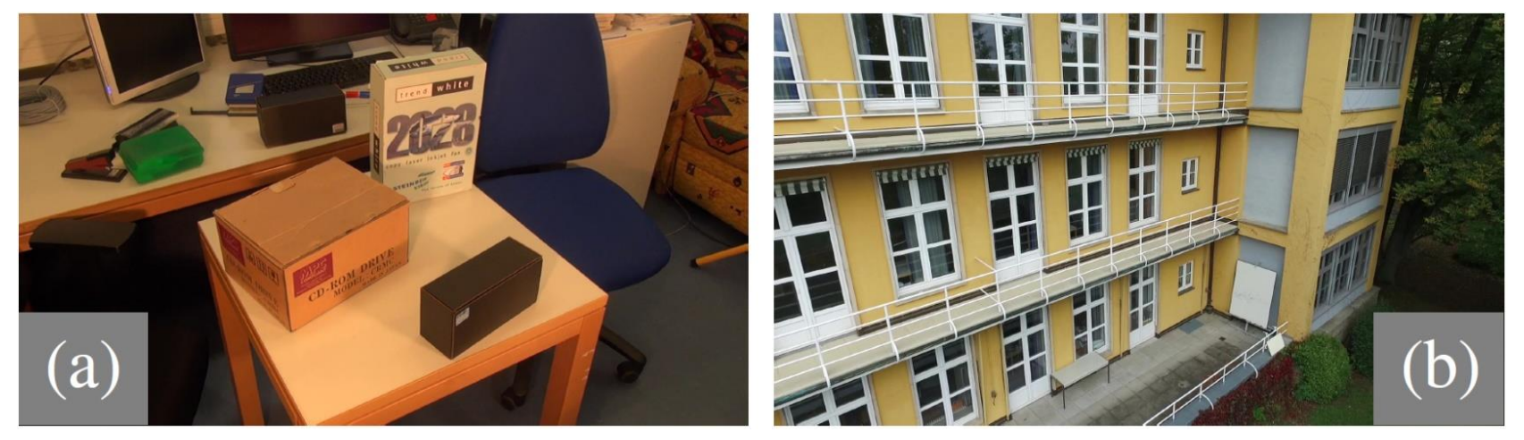

Fig. 4. Sampled images of real-world scenes.

As shown in Fig. 5, our reconstructed point clouds are significantly denser than BAI's, cleaner than KIM's, and smoother than WEI's. Furthermore, the surface discontinuities of our results align best with the object silhouettes. Especially, the numbers and letters in the Box scene, as well as the railings and window frames in the Building scene are all recognizable.

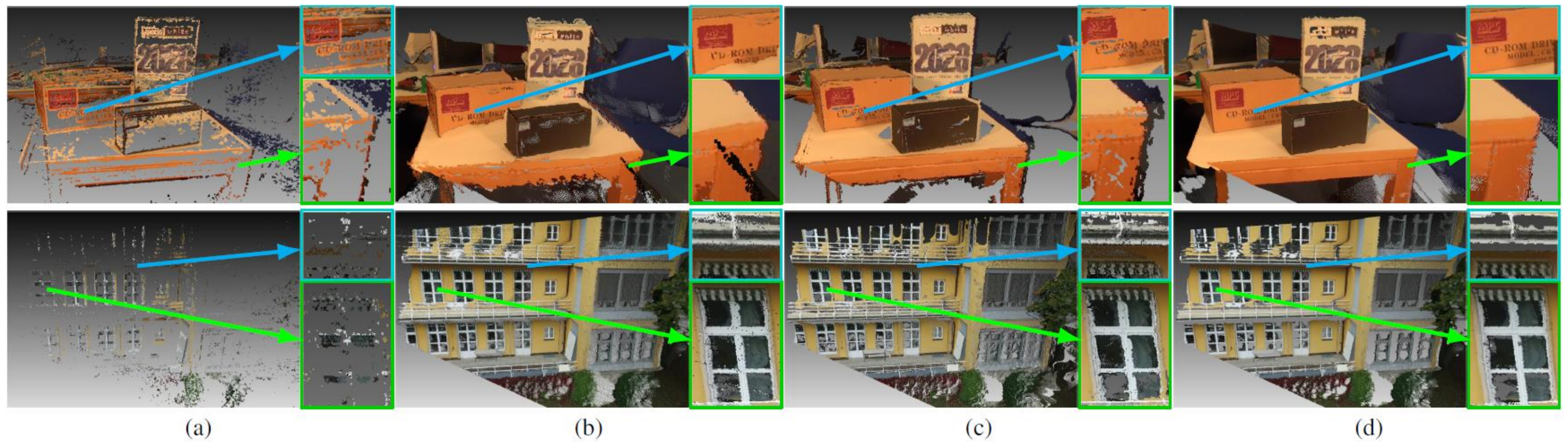

Fig. 5. Point clouds of the scenes in Fig. 4 reconstructed by (a) BAI, (b) KIM, (c) WEI, and (d) ours.

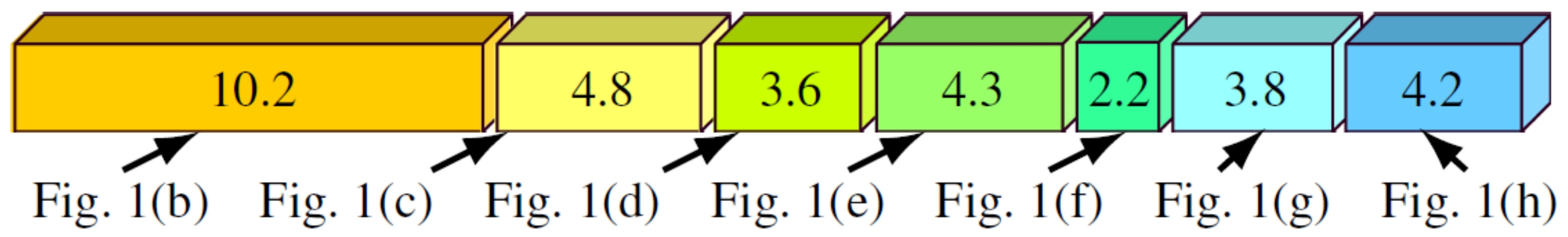

Fig. 6. Average runtime (sec.) for calculating individual immediate results in Fig. 1. Edge depth estimation takes the most time but is beyond the scope of this paper. Even so, the total runtime of 33.1 secs. is still less than those of BAI (71.8 secs.), KIM (122.2 secs.), and WEI (34.1 secs.).

References

[1] C. Bailer, M. Finckh, H. P. A. Lensch, "Scale robust multi view stereo," in *ECCV*, 2012.

[2] C. Kim, H. Zimmer, Y. Pritch, *et al.*, "Scene reconstruction from high spatioangular resolution light fields," *ACM Trans. Graph.* 32(73), 2013.

[3] J. Wei, B. Resch, H. P. A. Lensch, "Dense and occlusion-robust multi-view stereo for unstructured videos," in *Conf. Comput. and Robot Vis.*, 2016.