Omnidirectional 3D Reconstruction in Augmented Manhattan Worlds

Miriam Schönbein[1] and Andreas Geiger[2]

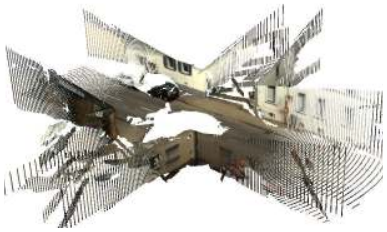
***Abstract*— This paper proposes a method for high-quality omnidirectional 3D reconstruction of augmented Manhattan worlds from catadioptric stereo video sequences. In contrast to existing works we do not rely on constructing virtual perspective views, but instead propose to optimize depth jointly in a unified omnidirectional space. Furthermore, we show that plane-based prior models can be applied even though planes in 3D do not project to planes in the omnidirectional domain. Towards this goal, we propose an omnidirectional slanted-plane Markov random field model which relies on plane hypotheses extracted using a novel voting scheme for 3D planes in omnidirectional space. To quantitatively evaluate our method we introduce a dataset which we have captured using our autonomous driving platform AnnieWAY which we equipped with two horizontally aligned catadioptric cameras and a Velodyne HDL-64E laser scanner for precise ground truth depth measurements. As evidenced by our experiments, the proposed method clearly benefits from the unified view and significantly outperforms existing stereo matching techniques both quantitatively and qualitatively. Furthermore, our method is able to reduce noise and the obtained depth maps can be represented very compactly by a small number of image segments and plane parameters.**

I. Introduction

3D perception is an important prerequisite for many tasks in robotics. For instance, consider self-driving vehicles [1] which need to accurately sense their environment in order to plan the next maneuver. Clearly, a 360° field of view is desirable. During the DARPA Urban Challenge [2] laser-based solutions have been popularized for that purpose. However, they provide only very sparse point clouds or are extremely expensive like the Velodyne HDL-64E. Furthermore, they suffer from rolling shutter effects and a separate video sensor is required to provide color information for each laser point. Instead, in this paper we advocate the use of catadioptric cameras [3], [4] for 3D reconstruction. Combining a traditional (perspective) camera with a mirror coated surface, they are cheap to produce and provide a 360° view of the scene which, in contrast to fisheye cameras can be parameterized by the specific choice of the mirror shape. Our setup is illustrated in Fig. 1(a).

We tackle the problem of reconstructing the static parts of 3D scenes which follow the augmented manhattan world assumption [5], i.e., scenes which can be described by vertical and horizontal planes in 3D. Note that this assumption does not require vertical planes to be orthogonal with respect to each other as in [6], [7], but only with respect to the horizontal planes. As illustrated in Fig. 1(b), many urban scenes closely follow this assumption and also indoor scenes are often composed of mainly horizontal or vertical surfaces [8], [9]. In this work, we show that incorporating such prior knowledge into the model can greatly benefit 3D reconstruction, in particular when dealing with omnidirectional images that often suffer from blur and low contrast.

(a) Stereo Camera Setup



(b) 3D Reconstruction

Fig. 1. **Omnidirectional 3D Reconstruction.** Figure (a) illustrates our catadioptric stereo camera setup. Figure (b) shows the result of the omnidirectional 3D reconstruction obtained by our method.

While planarity priors for stereo matching have been proposed in the context of traditional perspective cameras [10], [11], 3D planes do not project to planes in omnidirectional space, thereby preventing the use of classical prior models. We tackle this problem by proposing a slanted surface Markov random field (MRF) model based on superpixels in a virtual omnidirectional view. We start by spherically rectifying adjacent camera views and obtain initial depth estimates by matching each pair of views. Next, we aggregate all depth measurements in one common omnidirectional space and propose a Hough voting scheme which yields the set of dominant 3D planes in the scene. Subsequently, each 3D plane hypothesis is mapped to a non-linear surface in the omnidirectional image space, from which we compute potentials for all superpixels in the image. Plane optimization is formulated as a discrete labeling problem and carried out using loopy belief propagation which amounts to finding the best plane hypothesis for each superpixel under the assumption that nearby superpixels are likely belonging to the same surface.

Furthermore, we introduce a novel dataset of 152 diverse and challenging urban scenes for which we provide omnidirectional imagery as well as laser-based ground truth depth maps. We quantitatively show that our model outperforms state-of-the-art stereo matching techniques [12], [10] which have demonstrated superior performance in related evaluations such as the KITTI stereo benchmark [13]. We also show that our results are qualitatively more pleasing as

[1]Miriam Schönbein is with the Institute of Measurement and Control Systems, Karlsruhe Institute of Technology, 76131 Karlsruhe, Germany. miriam.schoenbein@kit.edu

[2]Andreas Geiger is with Max Planck Institute for Intelligent Systems, Perceiving Systems Department, 72076 Tübingen, Germany. andreas.geiger@tue.mpg.de

they are less susceptible to noise and allow for identifying the dominant planes which can be useful input information to subsequent higher-level reasoning stages such as scene understanding [14]. Our code, dataset and ground truth depth maps are publicly available[1].

II. Related Work

While there exists a large body of literature on omnidirectional camera calibration [15], [16], [17], [18], localization [19], [20] and sparse structure-from-motion / SLAM [21], [22], [23], surprisingly little research has been carried out towards dense 3D reconstruction with catadioptric cameras.

In [24], Svoboda and Pajdla investigate the epipolar geometry of central catadioptric systems. They show that the epipolar lines correspond to general conics in the omnidirectional image which reduce to radial lines for vertically aligned catadioptric cameras. As the latter allows for simple rectification, several methods take advantage of this setup by either mounting two catadioptric cameras on top of each other [25] or by using a double mirror design which allows for stereo matching with a single camera only [26]. Unfortunately, this configuration has a fixed and short baseline and only allows for accurate reconstructions in the very close range.

For general camera motion, [27], [28] propose to reproject the omnidirectional image to a panoramic image on a virtual cylinder. Stereo correspondences are established by searching along sinosoidal shaped epipolar curves [27], [29], [30]. Gonzalez and Lacroix [28] overcome this problem by rectifying the epipolar curves in panoramic images to straight lines. Similarly, Geyer and Daniilidis [31] present a conformal rectification method for parabolic images by mapping from bipolar coordinates to a rectangular grid. In this paper, we take advantage of spherical rectification [32], [33], [34] which is more flexible, can handle the existence of more than one epipole and does not depend on a particular projection model.

Towards dense 3D reconstruction, Arican and Frossard [34] obtain disparity maps from two omnidirectional views by optimizing a pixel-wise energy using graph cuts similar to the work of Fleck et al. [30]. Lhuiller [35] reconstructs the scene from three consecutive omnidirectional images which are projected onto the six faces of a virtual cube in order to allow for traditional stereo matching techniques. The local results are fused into a global model by selecting the most reliable viewpoints for each scene point and merging the 3D points using their median. This approach has been extended in [36] towards reconstruction of larger models from video sequences.

In contrast to the presented works that either consider 3D reconstruction from only two views or fuse depth maps in a rather ad-hoc manner, here we present a direct approach to 360° depth map optimization based on disparity estimates from two temporally and two spatially adjacent omnidirectional views. Note that due to the diverse spatial distribution of baselines this setup eliminates depth 'blind spots' which occur when reconstructing from two views only. Furthermore, we show how planarity priors can be incorporated directly in the omnidirectional domain, leading to clean low-noise 3D reconstructions. Accommodating the fact of limited public omnidirectional datasets, we contribute our data and the corresponding 3D ground truth to the robotics community.

[1]http://www.mrt.kit.edu/software/

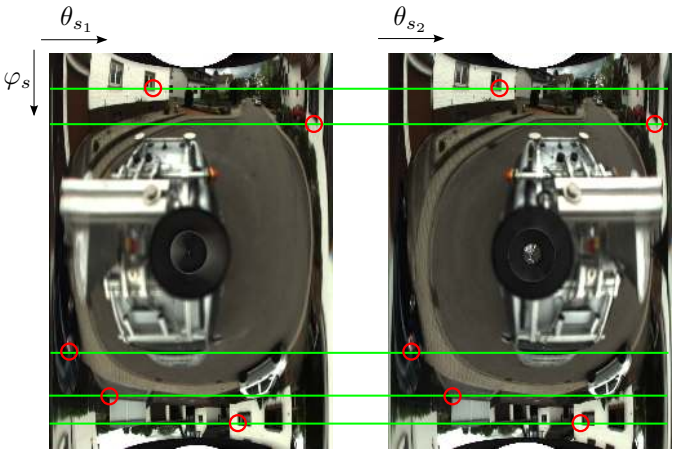


Fig. 2. **Rectified Catadioptric Stereo Pair.** The horizontal lines depict points with the same azimuth angle φ_s in the left and right image.

III. Omnidirectional Multi-View Stereo

To ease the formulation of the 3D reconstruction problem we first compute a virtual 360° disparity image from four omnidirectional views captured by a catadioptric stereo pair at two consecutive time steps. Through combination of depth information in one unified view we enable efficient inference and overcome the problem of blind spots near the epipoles [37] or occluded regions in some of the images. We calibrate our omnidirectional stereo camera rig using the method of [18] which optimizes for the best single viewpoint approximation even in cases where the cameras are slightly non-central, e.g., due to inaccuracies in the manufacturing process. Next, we estimate camera motion between two consecutive frames. We rectify temporal and spatial adjacent omnidirectional input pairs and combine their disparity maps in a single unified 360° inverse depth image which forms the basis for the plane-based inference discussed in Sec. IV. We discuss these steps in the following.

A. Motion Estimation

To estimate motion between two consecutive frames captured by the catadioptric stereo camera rig, we match sparse features between all views of both consecutive stereo pairs. We employ the FAST corner detector [38] in combination with the BRIEF descriptor [39], which empirically led to the best results for our images which suffer from blur at the image boundaries and incidental noise due to small scratches in the mirror surface. In practice, we obtain around 1300 correspondences in temporal and spatial direction. Using the extrinsic calibration of the stereo camera rig, we triangulate

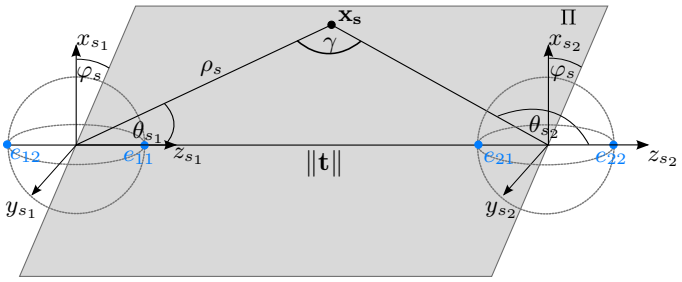
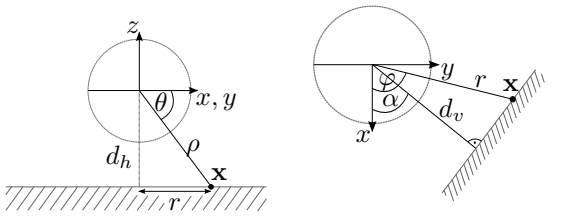


Fig. 3. **Spherical rectification.** After applying the rectifying rotation a 3D point $\mathbf{x}$ lies on the plane Π with the same azimuth angle φ_s in both (rotated) spherical coordinate systems. The rotated coordinate system depends on the relative position of both cameras determined by extrinsic calibration (stereo) or motion estimation (motion stereo), respectively.



(a) Horizontal Plane (Side View) (b) Vertical Plane (Top View)

Fig. 4. **Plane hypotheses.** This figure shows the relationship between a point $\mathbf{x}$ described by the spherical parameters φ, θ and its depth r, and the plane parameters d_h, d_v and α for horizontal and vertical planes in the coordinate system of the virtual camera.

feature points in the previous frame $t-1$ and estimate motion by minimizing the reprojection error with respect to the observations in the current frame t similar to the StereoScan system [40] for perspective cameras. Robustness against outliers is achieved by using RANSAC with 150 iterations. To balance the spatial distribution of feature points we employ bucketing using 16 cells with a maximum of 12 features per cell.

B. Rectification

To allow for efficient stereo matching, we rectify all four omnidirectional stereo pairs using spherical rectification similar to [32], [34]. This is illustrated in Fig. 3. We rotate an image pair in the spherical domain such that the epipoles coincide with the coordinate poles (z-axis). The remaining degree of freedom is chosen to minimize the relative rotation of the y-axis with respect to the camera coordinate system. Thus epipolar great circles coincide with the longitudes and disparity estimation reduces to a onedimensional search problem with constant azimuth angle φ_s. Fig. 2 depicts the result of the spherical rectification process. For further details we refer the reader to the appendix.

C. 360° Disparity Image

Given the rectified image pairs, we obtain disparity maps using semi-global matching [12] which has shown excellent performance in state-of-the-art perspective stereo benchmarks such as the KITTI stereo evaluation [13]. In the spherical domain the angular disparity $\gamma = \theta_{s_2} - \theta_{s_1}$ is defined as the difference between the angles θ_{s_1} and θ_{s_2} of the two viewing rays imaging the same 3D world point $\mathbf{x}_s$. The depth ρ_s of $\mathbf{x}_s$ is then given as

$$\rho_s = \frac{\|\mathbf{t}\| \cdot \sin\theta_{s_2}}{\sin\gamma} \tag{1}$$

where $\|\mathbf{t}\|$ denotes the baseline between the cameras. Due to the fact, that the images are highly distorted near the epipoles (leading to increased reconstruction error) as well as occlusions by the recording platform itself, we extract stereo depth estimates only from the front- (120°) and backward (120°) parts of the ego-vehicle while motion disparity is extracted only from the corresponding side (each 120°). After triangulating all points in the four spherical image pairs $I_{s_j}(\theta_s, \varphi_s)$ (with j denoting the image number), we project them into a new virtual 360° image $I(\varphi, \theta)$ with the camera coordinate system located in the center of the four original views. We choose the center as origin for the virtual coordinate system to minimize the relative displacement of all reflected rays. Furthermore, all points are rotated such that the x-y-plane of the new coordinate system is parallel to the groundplane. We estimate this transformation by computing the dominant plane below the camera using RANSAC plane fitting of the 3D points. Depth values for overlapping regions are merged by computing the mean value for each of these pixels. Fig. 5 (left) illustrates the resulting virtual 360° intensity image $I(\varphi, \theta)$ and inverse depth image $D(\varphi, \theta)$, where we define inverse depth by $D = 1/r$ with $r = \sqrt{x^2 + y^2}$ independent of the z-component of each 3D point to ease the representation of planes as will be discussed in Sec. IV. Note that working with inverse depth instead of depth implicitly accounts for the error characteristics of the underlying stereo measurements.

IV. Slanted-Plane MRF

For efficient inference and to propagate information over larger distances, we first partition the image into ~ 1000 superpixels using the StereoSLIC algorithm [10] applied to the 360° inverse depth image $D(\varphi, \theta)$ from the previous section. Next, we extract the set of dominant plane hypotheses for each scene and the problem of finding the best plane per superpixel is cast as a discrete labeling problem. The estimation of the plane hypotheses and the energy formulation are presented in the following.

A. Plane Hypotheses

Based on the fact that our coordinate system is parallel to the groundplane (x-y), we are able to describe vertical planes using two variables (angle α and distance d_v) and horizontal planes with a single variable only (distance d_h) as illustrated in Fig. 4. Since the depth r is independent from the z-component, the relationship between a 3D point $\mathbf{x}$ and the distance of a plane passing through $\mathbf{x}$ is given by

$$d_h(r, \theta) = \frac{r}{\tan\theta} \tag{2}$$

$$d_v(r, \varphi, \alpha) = r \cdot \cos(\varphi - \alpha) \tag{3}$$

where the variables denote angles and distances as defined in Fig. 4. This suggests a simple hough voting scheme: We accumulate the votes of all pixels in the virtual omnidirectional image in a 1-dimensional horizontal plane accumulator array $H(d_h)$ and in a 2-dimensional vertical plane accumulator array $H(d_v, \alpha)$ as illustrated in Fig. 5 (middle). To make the votes more discriminative, we disambiguate pixels belonging to horizontal and vertical surfaces by casting each vote with an additional weight which corresponds to the likelihood of a pixel belonging to a horizontal (or vertical) surface. This likelihood is modeled by logistic regression using the vertical inverse depth gradients as input. We estimate the parameters of the sigmoid function using a held out training set for which all horizontal and vertical surfaces have been manually labeled. The maxima of the voting accumulators $H(d_h)$ and $H(d_v, \alpha)$ are computed using an efficient non-maxima suppression implementation [41] which we have modified to handle cyclic image panoramas.

B. Energy Formulation

Given the plane hypotheses from the previous section, we formulate the problem of assigning each superpixel to one of the planes as a discrete energy minimization problem. More formally, let $\mathcal{S} = \{s_1, \ldots, s_M\}$ denote the variables of interest, each corresponding to one of the superpixels, where s takes a discrete plane index $s \in \{1, \ldots, N\}$ as value. Here, M denotes the total number of superpixels in the image and N is the number of plane hypotheses. We define the energy function to be minimized as

$$\Psi(\mathcal{S}) = \sum_{s \in \mathcal{S}} \left[\psi_{u_1}(s) + \psi_{u_2}(s)\right] + \sum_{(s_1, s_2) \in \mathcal{N}_\mathcal{S}} \psi_p(s_1, s_2) \quad (4)$$

with unary terms ψ_u and pairwise terms ψ_p where $\mathcal{N}_\mathcal{S}$ denotes the set of neighboring superpixels, i.e., all superpixels that share a common boundary.

The first unary term models the inverse depth fidelity

$$\psi_{u_1}(s) = w_{u_1} a(s) \sum_{p \in \mathcal{P}_s} \left[\rho_u(\hat{D}(p, s) - D(p))\right] \quad (5)$$

with weight parameter w_{u_1}. Here, $\hat{D}(p, s)$ is the inverse depth at pixel $p = (\varphi, \theta)^T$ predicted from the plane with index s, $D(p)$ is the inverse depth estimate at pixel p (see Sec. III-C) and $\rho_u(x) = \min(|x|, \tau_u)$ is a robust l_1 penalty function with truncation parameter τ_u. Furthermore, $\mathcal{P}_s$ denotes the set of all pixels with valid inverse depth hypothesis $D(p)$ which are covered by superpixel s and $a(s) \in [0, 1]$ is a function that predicts the accuracy of the inverse depth map D averaged over superpixel s from training data. The latter has been introduced as we found the reliability of SGM to correlate strongly with image blur and hence also image location when dealing with omnidirectional images. In practice, we take $a(s)$ as the average ratio of correctly predicted depth values computed from a held-out training set.

The second unary term models the prior probability for surfaces to be horizontal or vertical and is given by

$$\psi_{u_2}(s) = w_{u_2} \times \begin{cases} 2\,p_h(s) - 1 & \text{if } s \in \mathcal{H} \\ 1 - 2\,p_h(s) & \text{otherwise} \end{cases} \quad (6)$$

where $\mathcal{H}$ is the set of horizontal planes and

$$p_h(s) = \frac{1}{|\mathcal{P}_s|} \sum_{p \in \mathcal{P}_s} p'_h(p) \in [0, 1] \quad (7)$$

is the prior probability of superpixel s being horizontal. Here, $p'_h(p)$ is simply the probability of *pixel* p being horizontal which we compute from our held-out training set augmented with manually labeled polygons of vertical and horizontal surfaces. Note how (6) assigns a positive score to plane hypotheses that agree with the expected plane type and negative scores otherwise.

Our pairwise model encourages neighboring superpixels to agree at their boundaries

$$\psi_p(s_1, s_2) = w_p \sum_{p \in \mathcal{B}_{s_1, s_2}} \rho_p(\hat{D}(p, s_1) - \hat{D}(p, s_2)) \quad (8)$$

where w_p is a smoothness parameter and $\mathcal{B}_{s_1, s_2}$ is the set of boundary pixels that are shared between s_1 and s_2. Similar to the depth fidelity term, we take $\rho_u(x) = \min(|x|, \tau_p)$ as the robust l_1 penalty with truncation parameter τ_p.

C. Learning and Inference

For inferring $\mathcal{S}$ we make use of min-sum loopy belief propagation to approximately minimize the energy specified in (4). The parameters of our model are estimated from a separate training set consisting of 80 images. As (4) depends nonlinearly on τ_u and τ_p, traditional CRF learning algorithms [42] are not feasible and we resort to Bayesian optimization [43] for estimating the parameters, yielding $w_{u_1} = 1.2$, $w_{u_2} = 1.0$, $w_p = 1.0$, $\tau_u = 0.05$ and $\tau_p = 0.08$.

V. Evaluation

We evaluate our approach using stereo sequences captured with our autonomous driving platform AnnieWAY. We equipped the vehicle with two horizontally aligned hypercatadioptric cameras on top of the roof of the vehicle, a high-precision GPS/IMU system that delivers groundtruth motion and a Velodyne laser scanner that provides 360° laser scans with a vertical resolution of 64 laser beams.

A. Ground truth

We use the Velodyne laser scanner as reference sensor for our quantitative evaluation. As we focus on static scenes only, we are able to accumulate the laser point clouds (+/- 5 frames) using ICP point-to-plane fitting which yields relatively dense ground truth depth maps (see Fig. 6(a) (top) for an illustration). The calibration between the catadioptric camera and the Velodyne laser scanner is obtained by minimizing the reprojection error in the image from manually selected correspondences. To evaluate the quality of depth information depending on surface inclination, we also labeled all horizontal and vertical planes and obtain the

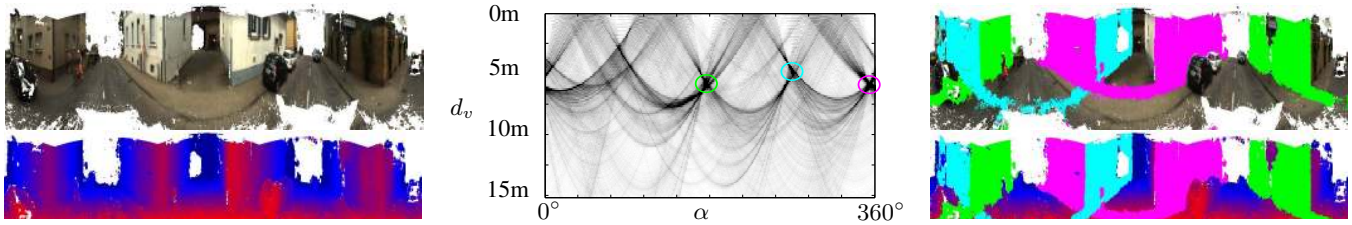


Fig. 5. **Plane Hypotheses.** This figure shows the virtual omnidirectional intensity image (left top) and the corresponding false color depth map from SGBM (left bottom), the Hough space for vertical planes (middle) and the intensity and inverse depth image with three randomly selected planes (right) corresponding to the colored maxima in the Hough space. Note how the cyan maximum describes a plane that is closer to the camera center (smaller d_v) than the planes corresponding to the green and purple maxima. This can also be verified by looking at the plane visualizations on the right.

plane parameters using the Hough transformation presented in Sec. IV-A with the ground truth depth maps as input (in contrast to the estimated ones used in our method). Our dataset comprises 80 training and 72 test scenes in total.

B. Quantitative Results

We evaluate the proposed method against state-of-the-art stereo vision algorithms. Our baselines include simple Block Matching (BM), Semi-Global Matching (SGBM) [12] (in both cases we made use of the OpenCV implementations), as well as the more recently developed StereoSLIC algorithm [10]. To investigate the importance of the proposed plane-based prior, we also implement a winner takes all (WTA) plane selection strategy, which selects the best plane independently for each superpixel. Note that this corresponds to minimizing (4) while ignoring pairwise potentials $\psi_p(s_1, s_2)$ and the horizontal prior $\psi_{u_2}(s)$.

We compute the inverse depth error $e = |D_{gt} - D_{est}|$ for every pixel for which groundtruth is available. To guarantee a fair comparison, we fill in missing values in the resulting inverse depth images using background interpolation [12], [13]. We report the mean number of bad pixels and the mean end-point error averaged over the full test set. A pixel which has an inverse depth error e larger than $0.05\ 1/m$ is regarded as a bad pixel. Tab. I shows the mean percentage of bad pixels and the mean end-point error for all algorithms averaged over all 72 test images. The first row depicts the errors for all pixels where depth ground truth is available, while the other rows consider planar regions only. For WTA we vary the threshold of our non-maxima suppression stage between 50 and 500 (WTA 50 / WTA 500 in Tab. I), yielding about 5 to 150 planes on average. For our method, we set this threshold to a constant value of 150.

Our experiments show that the proposed method significantly outperforms the baselines. The difference is especially pronounced for horizontal planes, but our method also decreases the number of bad pixels for vertical planes with respect to all baseline methods.

C. Qualitative Results

Fig. 6 and 7 depict the inverse depth images for the analyzed algorithms (b-f) and the inverse depth groundtruth obtained from the Velodyne laser scanner (a). Colors represent distance, where green is close and blue denotes distant points. Alongside, we show the 3D reconstructions obtained when reprojecting all pixels of the corresponding inverse depth maps back into 3D (b-f). Note how our algorithm is able to produce much cleaner depth images and smoother 3D reconstructions. A random selection of challenging 3D scenes reconstructed using our method is given in Fig. 8.

VI. Conclusions and Future Work

In this paper we presented a method for high-quality omnidirectional 3D reconstruction from a single virtual inverse depth image. We showed how efficient inference with plane-based prior models is possible and leads to clean and easy to interpret depth maps that outperform state-of-the art depth estimation techniques in terms of 3D reconstruction error. In the future, we plan to investigate possible extensions towards integrating depth information from more than four views to allow for example for urban reconstructions at larger scales.

Appendix

This appendix provides details of the spherical rectification outlined in Sec. III-B. For clarity we only illustrate the process for the first (reference) camera. The mapping for the second camera is obtained in a similar manner.

Let $I_o(u, v)$ denote the omnidirectional input image with pixel coordinates $(u, v)^T$ and let $I_s(\theta_s, \varphi_s)$ denote the rectified spherical image which depends on the azimuth angle $\varphi_s \in [0, 2\pi]$ and inclination angle $\theta_s \in [0, \pi]$ as illustrated in Fig. 3. We obtain

$$\varphi_s = \arctan \frac{y_r}{x_r} \qquad \theta_s = \arctan \frac{\sqrt{x_r^2 + y_r^2}}{z_r} \tag{9}$$

where $(x_r, y_r, z_r)^T = \mathbf{R} \cdot (x, y, z)^T$, $\mathbf{R}$ is a rotation matrix and the ray $(x, y, z)^T$ corresponds to pixel $(u, v)^T$ in $I_o(u, v)$. The rectifying rotation matrix $\mathbf{R}$ is computed such that the epipoles coincide with the coordinate poles, i.e., all epipoles lie on the line connecting both camera centers. This is achieved by letting

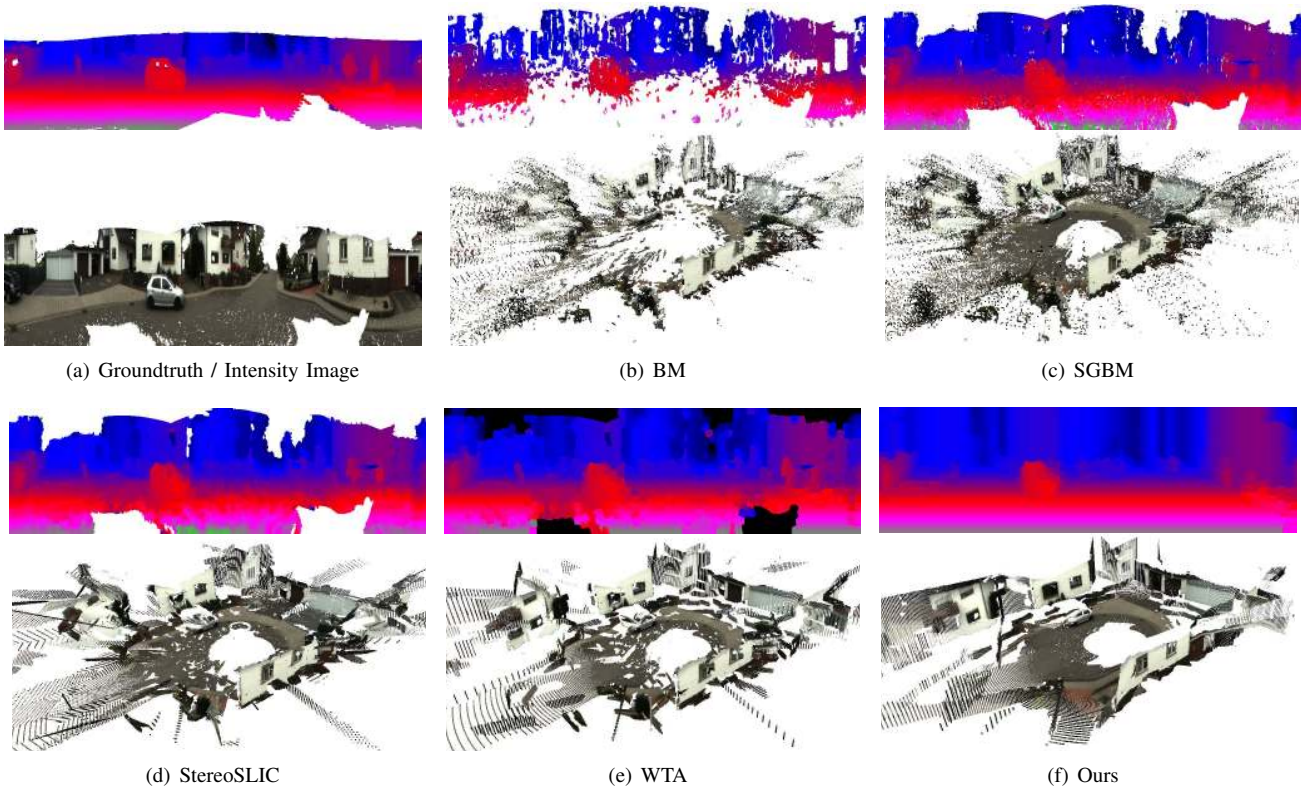
$$\begin{aligned} \mathbf{R} &= [\mathbf{r}_1, \mathbf{r}_2, \mathbf{e}_{11}] \\ \mathbf{r}_2 &= \mathbf{y}_o - (\mathbf{e}_{11}^T \mathbf{y}_0)\, \mathbf{e}_{11} \\ \mathbf{r}_1 &= \mathbf{r}_2 \times \mathbf{r}_3 \end{aligned}$$

where $\mathbf{y}_o$ denotes the y-axis of the original omnidirectional camera system (before rotation) and $\mathbf{e}_{11}$ is the first epipolar point as illustrated in Fig. 3. Note that this definition removes

Bad Pixels (%)	SGBM	BM	StereoSLIC	WTA 50	WTA 100	WTA 150	WTA 200	WTA 300	WTA 500	Ours
All Pixel	11.89	9.52	8.95	11.62	11.63	11.59	11.62	12.63	14.66	**4.04**
All Planes	13.41	7.27	9.50	13.22	13.16	12.85	12.28	11.96	11.98	**1.24**
Horizontal Planes	17.45	6.75	12.24	17.48	17.40	17.04	16.33	15.29	13.28	**1.03**
Vertical Planes	2.52	5.81	1.85	2.11	2.10	2.20	2.33	6.64	14.10	**1.51**
Mean Error ($1/r$)	SGBM	BM	StereoSLIC	WTA 50	WTA 100	WTA 150	WTA 200	WTA 300	WTA 500	Ours
All Pixel	0.026	0.022	0.021	0.029	0.029	0.029	0.029	0.030	0.031	**0.013**
All Planes	0.029	0.022	0.022	0.033	0.033	0.032	0.031	0.030	0.030	**0.009**
Horizontal Planes	0.034	0.023	0.026	0.038	0.038	0.037	0.036	0.034	0.032	**0.010**
Vertical Planes	**0.008**	0.013	**0.008**	**0.008**	**0.008**	0.009	0.009	0.016	0.023	**0.008**

TABLE I

Quantitative Analysis. THIS TABLE SHOWS THE MEAN PERCENTAGE OF BAD PIXELS AND THE MEAN INVERSE DEPTH ERROR FOR ALL BASELINES AND THE PROPOSED METHOD AVERAGED OVER ALL 72 TEST IMAGES. THE FIRST ROW DEPICTS THE ERRORS FOR ALL PIXELS WHERE DEPTH GROUND TRUTH IS AVAILABLE, WHILE THE OTHER ROWS CONSIDER PLANAR REGIONS (OF A SPECIFIC TYPE) ONLY.



(a) Groundtruth / Intensity Image (b) BM (c) SGBM

(d) StereoSLIC (e) WTA (f) Ours

Fig. 6. **Inverse Depth Maps and 3D Reconstructions.** The figures show the inverse depth images and the resulting 3D reconstruction for the same scene for the baseline algorithms (BM, SGM, StereoSLIC), for the best WTA result with threshold 150 and our MRF based plane estimation.

the remaining degree of freedom by ensuring that the rotated y-axis is similar to the original one. The epipoles are obtained from the essential matrix $\mathbf{E} = [\mathbf{t}]_\times \mathbf{R}$ which is specified by the rigid motion $[\mathbf{R}|\mathbf{t}]$ between both cameras.

REFERENCES

[1] A. Geiger, M. Lauer, F. Moosmann, B. Ranft, H. Rapp, C. Stiller, and J. Ziegler, "Team annieway's entry to the grand cooperative driving challenge 2011," *TITS*, 2012.

[2] M. Buehler, K. Iagnemma, and S. Singh, Eds., *The DARPA Urban Challenge*, ser. Advanced Robotics, vol. 56, 2009.

[3] S. Baker and S. K. Nayar, "A theory of single-viewpoint catadioptric image formation," *IJCV*, 1999.

[4] C. Geyer and K. Daniilidis, "A unifying theory for central panoramic systems and practical implications," in *ECCV*, 2000.

[5] G. Schindler and F. Dellaert, "Atlanta world: An expectation maximization framework for simultaneous low-level edge grouping and camera calibration in complex man-made environments," in *CVPR*, 2004.

[6] Y. Furukawa, B. Curless, S. M. Seitz, and R. Szeliski, "Manhattan-world stereo," in *CVPR*, 2009.

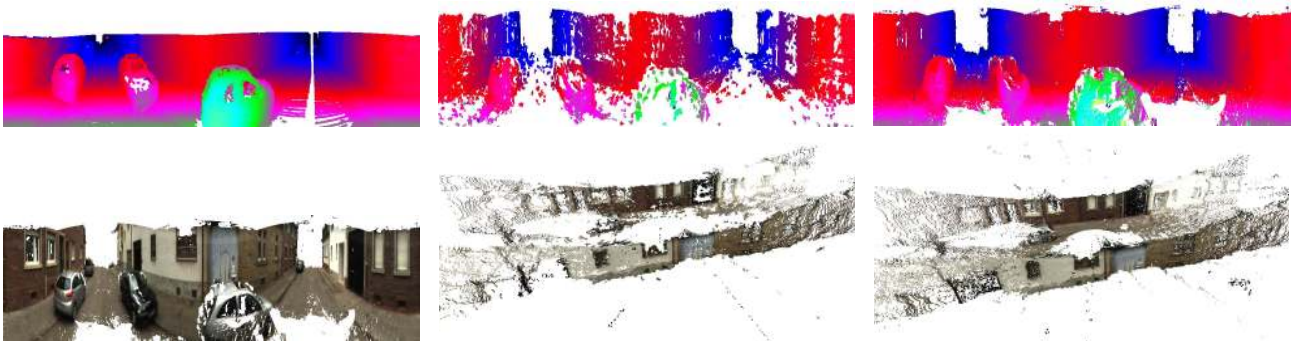
[7] ——, "Reconstructing building interiors from images," in *ICCV*, 2009.

[8] A. Schwing and R. Urtasun, "Efficient exact inference for 3d indoor scene understanding," in *ECCV*, 2012.

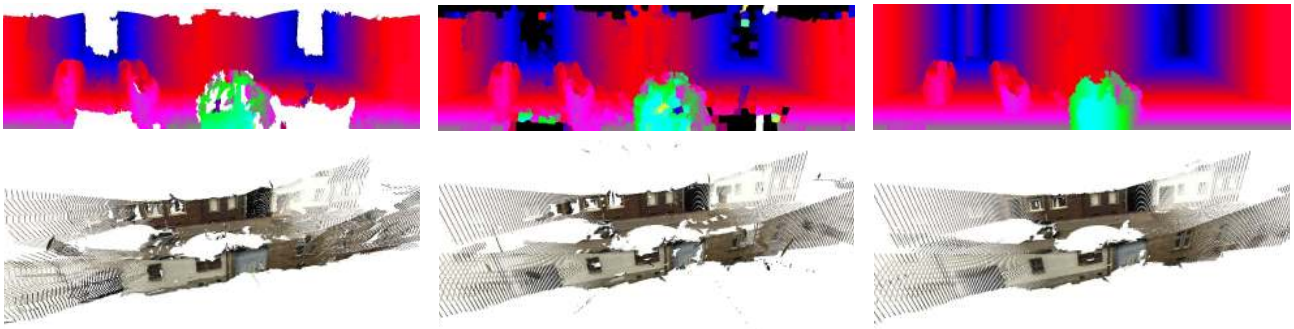
[9] B. Zeisl, C. Zach, and M. Pollefeys, "Stereo reconstruction of building interiors with a vertical structure prior," in *THREEDIMPVT*, 2011.

[10] K. Yamaguchi, D. McAllester, and R. Urtasun, "Robust monocular epipolar flow estimation," *CVPR*, 2013.

[11] B. Mičušík and J. Košecká, "Multi-view superpixel stereo in urban



(a) Groundtruth / Velodyne (b) BM (c) SGBM



(d) StereoSLIC (e) WTA (f) Ours

Fig. 7. **Inverse Depth Maps and 3D Reconstructions.** The figures show the inverse depth images and the resulting 3D reconstruction for the same scene for the baseline algorithms (BM, SGM, StereoSLIC), for the best WTA result with threshold 150 and our MRF based plane estimation.

environments," *IJCV*, 2010.

[12] H. Hirschmüller, "Stereo processing by semiglobal matching and mutual information," *PAMI*, 2008.

[13] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *CVPR*, 2012.

[14] A. Geiger, M. Lauer, C. Wojek, C. Stiller, and R. Urtasun, "3d traffic scene understanding from movable platforms," *PAMI*, 2014.

[15] D. Scaramuzza and A. Martinelli, "A toolbox for easily calibrating omnidirectional cameras," in *IROS*, 2006.

[16] C. Mei and P. Rives, "Single view point omnidirectional camera calibration from planar grids," in *ICRA*, 2007.

[17] L. Puig, J. Bermúdez, P. Sturm, and J. J. Guerrero, "Calibration of omnidirectional cameras in practice: A comparison of methods," *CVIU*, 2012.

[18] M. Schönbein, T. Strauss, and A. Geiger, "Calibrating and centering quasi-central catadioptric cameras," in *ICRA*, 2014.

[19] A. Murillo, J. Guerrero, and C. Sagues, "Surf features for efficient robot localization with omnidirectional images," in *ICRA*, 2007.

[20] C. Valgren and A. J. Lilienthal, "Sift, surf and seasons: Long-term outdoor localization using local features." in *EMCR*, 2007.

[21] A. Pagani and D. Stricker, "Structure from motion using full spherical panoramic cameras," in *ICCV Workshops*, 2011.

[22] J.-P. Tardif, Y. Pavlidis, and K. Daniilidis, "Monocular visual odometry in urban environments using an omnidirectional camera," in *IROS*, 2008.

[23] A. Rituerto, L. Puig, and J. J. Guerrero, "Visual slam with an omnidirectional camera," in *ICPR*, 2010.

[24] T. Svoboda and T. Pajdla, "Epipolar geometry for central catadioptric cameras," *IJCV*, 2002.

[25] J. Gluckman, S. K. Nayar, and K. J. Thoresz, "Real-time omnidirectional and panoramic stereo," in *In DARPA Image Understanding Workshop*, 1998.

[26] S. Yi and N. Ahuja, "An omnidirectional stereo vision system using a single camera," in *ICPR*, 2006.

[27] R. Bunschoten, B. J. A. Krse, and N. A. Vlassis, "Robust scene reconstruction from an omnidirectional vision system." *IEEE T. Robotics and Automation*, vol. 19, no. 2, pp. 351–357, 2003.

[28] J.-J. Gonzalez-Barbosa and S. Lacroix, "Fast dense panoramic stereovision." in *ICRA*, 2005.

[29] S. B. Kang and R. Szeliski, "3-d scene data recovery using omnidirectional multibaseline stereo," *IJCV*, 1995.

[30] S. Fleck, F. Busch, P. Biber, W. Strasser, and H. Andreasson, "Omnidirectional 3d modeling on a mobile robot using graph cuts," in *ICRA*, 2005.

[31] C. G. Kostas and K. Daniilidis, "Conformal rectification of omnidirectional stereo pairs," in *Omnivis*, 2003.

[32] J. Fujiki, A. Torii, and S. Akaho, "Epipolar geometry via rectification of spherical images," in *Computer Vision/Computer Graphics Collaboration Techniques*, 2007.

[33] S. Li, "Real-time spherical stereo," in *ICPR*, 2006.

[34] Z. Arican and P. Frossard, "Dense disparity estimation from omnidirectional images." in *AVSS*, 2007.

[35] M. Lhuillier, "Toward flexible 3d modeling using a catadioptric camera," in *CVPR*, 2007.

[36] S. Yu and M. Lhuillier, "Surface reconstruction of scenes using a catadioptric camera," in *Computer Vision/Computer Graphics Collaboration Techniques*, 2011.

[37] M. Schönbein, H. Rapp, and M. Lauer, "Panoramic 3d reconstruction with three catadioptric cameras," in *IAS*, 2013.

[38] E. Rosten and T. Drummond, "Machine learning for high-speed corner detection." Springer Berlin Heidelberg, 2006, pp. 430–443.

[39] M. Calonder, V. Lepetit, M. Ozuysal, T. Trzcinski, C. Strecha, and P. Fua, "Brief: Computing a local binary descriptor very fast," *PAMI*, 2012.

[40] A. Geiger, J. Ziegler, and C. Stiller, "StereoScan: Dense 3d reconstruction in real-time," in *IV*, 2011.

[41] A. Neubeck and L. V. Gool, "Efficient non-maximum suppression," in *ICPR*, 2006.

[42] T. Hazan and R. Urtasun, "A primal-dual message-passing algorithm for approximated large scale structured prediction," in *NIPS*, 2010.

[43] H. L. Jasper Snoek and R. P. Adams, "Practical bayesian optimization of machine learning algorithms," in *NIPS*, 2012.

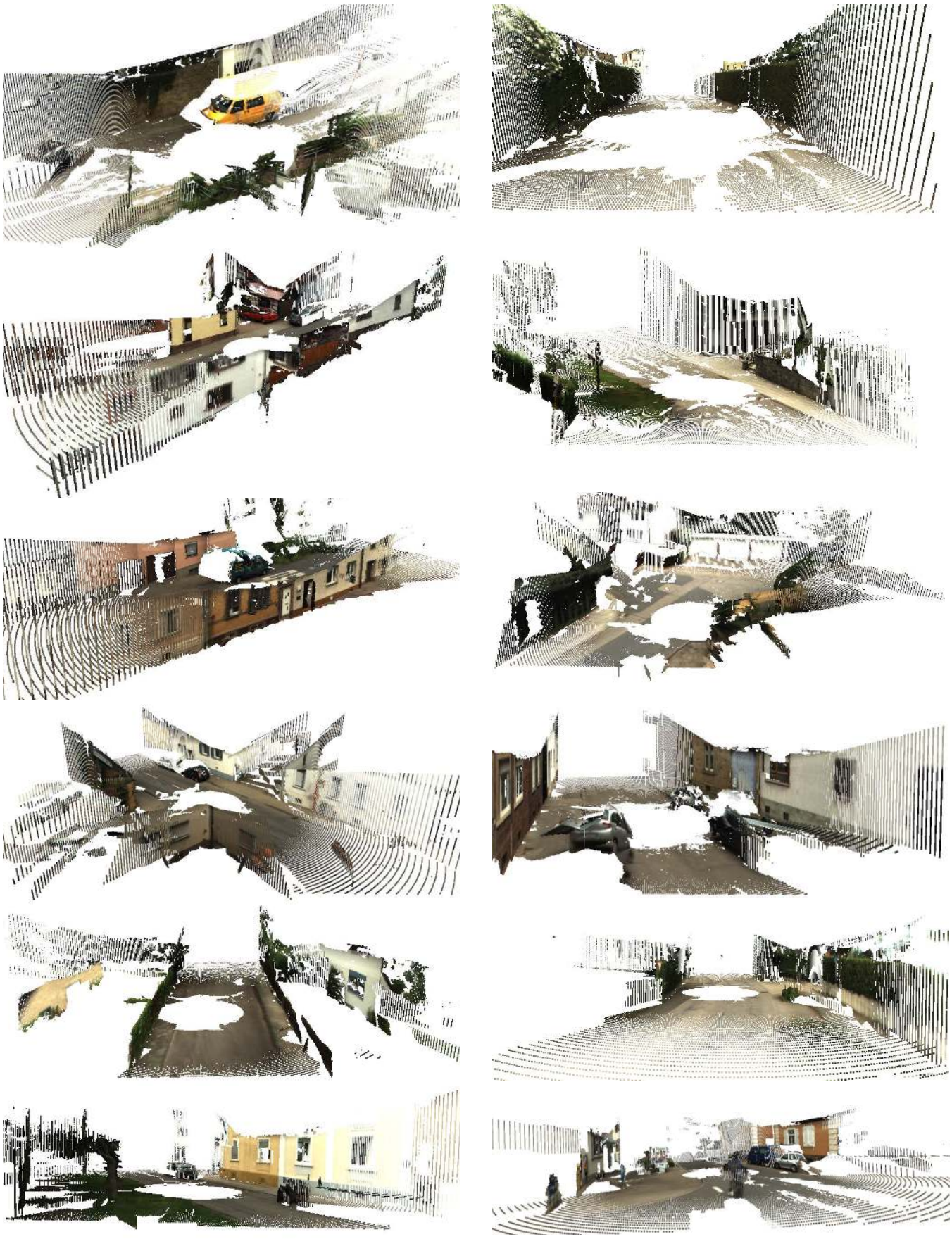


Fig. 8. **3D Reconstruction:** This figure shows 3D reconstructions for different urban scenarios obtained when reprojecting the inverse depth maps produced by our method into 3D. Note that the viewpoint of the rendered 3D point clouds deviates significantly from the viewpoint of the four cameras.