

Project AutoVision: Localization and 3D Scene Perception for an Autonomous Vehicle with a Multi-Camera System

Lionel Heng¹, Benjamin Choi¹, Zhaopeng Cui², Marcel Geppert², Sixing Hu³, Benson Kuan¹, Peidong Liu², Rang Nguyen³, Ye Chuan Yeo¹, Andreas Geiger⁴, Gim Hee Lee³, Marc Pollefeys^{2,5}, and Torsten Sattler⁶

Abstract—Project AutoVision aims to develop localization and 3D scene perception capabilities for a self-driving vehicle. Such capabilities will enable autonomous navigation in urban and rural environments, in day and night, and with cameras as the only exteroceptive sensors. The sensor suite employs many cameras for both 360-degree coverage and accurate multi-view stereo; the use of low-cost cameras keeps the cost of this sensor suite to a minimum. In addition, the project seeks to extend the operating envelope to include GNSS-less conditions which are typical for environments with tall buildings, foliage, and tunnels. Emphasis is placed on leveraging multi-view geometry and deep learning to enable the vehicle to localize and perceive in 3D space. This paper presents an overview of the project, and describes the sensor suite and current progress in the areas of calibration, localization, and perception.

I. INTRODUCTION

The three DARPA Grand Challenges in the last decade set off a wave of disruption in the automotive industry. With widespread belief that autonomous vehicles can revolutionize logistics and mobility, automakers and technology companies are racing with one another to put autonomous vehicles on the road within the next few years. LiDAR sensors are the primary sensing modality for a vast majority of autonomous vehicles; they generate highly accurate 3D point cloud data in both day and night, and enable localization and 3D scene perception at all times of the day. In contrast, cameras require sufficient ambient lighting, and do not directly provide 3D point cloud data. However, cameras yield high-resolution image data which better facilitates scene segmentation and understanding. In addition, we can leverage multi-view geometry techniques to infer depth data from multiple cameras albeit with lower accuracy than depth data from LiDAR sensors. Cameras can be fitted with either wide-field-of-view or fisheye lenses, giving them a significantly larger vertical field of view and higher vertical resolution compared to LiDAR sensors. In Project AutoVision, we choose to focus on cameras as the sole sensing modality for autonomous vehicles; we observe that much research remains to be done in realising robust visual localization and perception for autonomous vehicles.

Project AutoVision started in late 2016 with the goal to develop localization and 3D scene perception algorithms

for autonomous vehicles exclusively equipped with cameras. Project AutoVision is similar to, but differs from Project V-Charge [9, 31, 14] in the aspect that Project AutoVision extends the operating envelope from parking lots and garages to large-scale urban and rural environments with higher driving speeds and widely varying illumination conditions. Parallels can also be drawn between Project AutoVision and AutoX, both of which only rely on cameras for localization and perception. However, little is known about AutoX's localization and perception approaches due to commercial interests.

We aim to localize in both mapped and unmapped areas without relying on GNSS; we do not want to limit the vehicle's operation to mapped areas in GNSS-less conditions. In addition, we work towards real-time 3D mapping as a 3D geometric map can aid navigation of multi-level structures and higher-level scene perception tasks such as terrain analysis and 3D scene understanding. We follow the traditional approach [33, 3] of applying multi-view geometry to localization and 3D geometric mapping, and machine learning to cross-modal matching, object detection, and scene segmentation. On the other hand, with the advent of deep neural networks, people [2] have used end-to-end learning for vision-based autonomous vehicles but with limited success.

We make the following contributions:

- 1) Real-time visual-inertial odometry with a multi-camera system [25, 26].
- 2) Real-time GNSS-less visual localization in unmapped areas using geo-referenced satellite imagery and without GNSS, assuming that the initial global position and heading of the vehicle are known [19, 18].
- 3) Real-time GNSS-less visual localization with a multi-camera system in mapped areas using a geo-referenced sparse 3D map and without prior knowledge of the vehicle's global pose [11].
- 4) Real-time 3D dense mapping with a multi-fisheye-camera system [5].

In this paper, we briefly describe these contributions which have been integrated into a single working system. The reader can refer to our published work for more details on the algorithms and experimental results. In addition, the paper gives details of the hardware setup, the software architecture, and the automated methods used for calibrating the multi-sensor suite; such details are not found in our published work on individual localization and perception components.

¹DSO National Laboratories

²ETH Zürich

³National University of Singapore

⁴MPI-IS and University of Tübingen

⁵Microsoft, Switzerland

⁶Chalmers University of Technology, Sweden



Fig. 1. The AutoVision vehicle platform.

II. SYSTEM

In this section, we give an overview of the sensors on the AutoVision vehicle platform, and the software architecture that enables various software modules to work together to enable the vehicle to localize and perceive in 3D.

Our AutoVision vehicle platform is a Isuzu D-Max pickup truck which has been modified to include a drive-by-wire system for autonomous driving. Fig. 1 shows the vehicle platform while Fig. 2 shows a close-up view of the sensors on the vehicle roof. Four color cameras and twelve NIR cameras are fitted with 180° -field-of-view fisheye lenses and installed in a surround-view configuration on top of the vehicle. All cameras output 2-megapixel images at 30 Hz, and are set to automatic exposure mode so that they can adapt to changing lighting conditions. We only use 12-bit grayscale images from the NIR cameras as input to all localization and perception modules; color cameras are only used for visualization purposes. NIR cameras are more light-sensitive than color cameras, and can detect light in both the visible and NIR wavelengths. In addition, NIR cameras provide sharp and clean images unlike Bayer-encoded color images that suffer from demosaicing artefacts. We use NIR cameras in conjunction with NIR illuminators which can improve low-light imaging quality and whose illumination is invisible to and does not distract drivers on the road. Fig. 4 shows examples of images captured at approximately the same location in varying lighting conditions. Camera enclosures provide cameras with IP67 protection from the weather elements. Fig. 3 shows the camera layout on the vehicle roof. The front side of the vehicle has the highest number of NIR cameras; these 5 NIR cameras facilitate wide-baseline stereo, and in turn, long-range perception which is critical for autonomous vehicles moving forward at high speeds. We exploit the dominantly longitudinal movement of the vehicle by simulating a multi-baseline stereo system on each side of the vehicle and which consists of 2 actual NIR cameras and at least 1 virtual NIR camera.

A. Hardware

A dual-antenna GNSS/INS system with a tactical-grade IMU is installed in the vehicle. Data from this GNSS/INS



Fig. 2. A close-up of the sensor suite on the AutoVision vehicle platform.

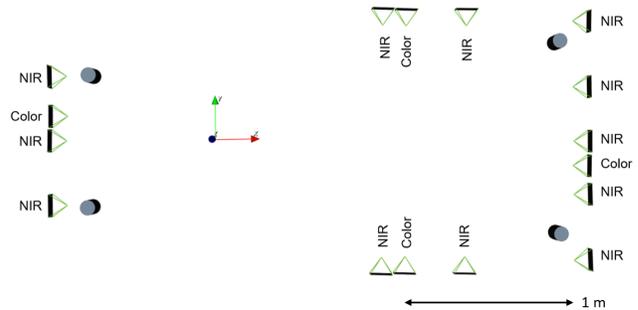


Fig. 3. The sensor layout on the AutoVision vehicle platform. Each green-colored frustum and grey-colored cylinder represent a camera and LiDAR sensor respectively. The origin of the three axes indicates the location of the IMU. The red-colored x -axis points towards the front of the vehicle while the green-colored y -axis points towards the left of the vehicle. The extrinsic transformation between each sensor and the IMU was estimated with automated calibration tools.

system is post-processed offline to yield centimeter-level ground-truth position and attitude data which is used to evaluate localization accuracy. A 3D LiDAR sensor is mounted at each of the four corners of the vehicle roof; fused point cloud data from all 4 LiDAR sensors is used to evaluate perception accuracy. All sensor data is hardware-timestamped to sub-microsecond precision. Such accurate time synchronization is made possible through the use of a time server. This time server synchronizes to the GNSS/INS system via PPS signals and NMEA data. In turn, LiDAR sensors and cameras time-synchronize to the time server via PPS/NMEA and PTP respectively. All sensors interface over a 10 GbE network switch with multiple industrial-grade computers equipped with GPUs. The 16-camera system yields 1.38 GB/s of image data which we record simultaneously to 4 solid-state drives (SSDs); we configure each computer to have two SSDs, and thus, two computers are required for data recording.

B. Software

Our software stack is based on the ROS 2 software framework, and runs on Windows 10. We use RTI Connex DDS for inter-process communications. Fig. 5 shows our software architecture; with the exception of the block representing sparse 3D map reconstruction which runs offline, each block represents a node which subscribes and publishes to topics. GNSS and IMU measurements are only used by sparse 3D



Fig. 4. Images captured from a front left camera on the AutoVision vehicle and in varying lighting conditions.

map reconstruction and visual-inertial odometry respectively. All nodes subscribe to image data.

III. CALIBRATION

An accurate calibration is an essential prerequisite for localization and perception to work well with a multi-camera system. Our calibration pipeline is automated and involves the following steps in order: intrinsic and extrinsic calibration of a multi-camera system, extrinsic calibration between a calibrated multi-camera system and a GNSS/INS system, and extrinsic calibration between a LiDAR sensor and a calibrated multi-camera system.

We do intrinsic and extrinsic calibration of the multi-camera system with the help of a fiducial target which is a grid of AprilTag markers [32] with known dimensions, and is shown in Fig. 6. As each AprilTag marker has a unique identifier, calibration will work even if multiple cameras observe different parts of the target. This versatility comes in handy when calibrating pairs of cameras at the vehicle’s corners and with perpendicular optical axes; it is difficult for such a pair of cameras to observe the entire target. For the intrinsic calibration, we can choose from multiple camera projection and distortion models. In this project, we use the unified projection model [12, 1] and the plumb bob distortion model [4]. We also perform a photometric calibration of each camera using the method described in [7]. This photometric calibration is useful for photometric-based matching between images captured with different exposure times, for example, in direct visual odometry and plane-sweeping stereo.

We obtain the extrinsic transformation between the calibrated multi-camera system and the GNSS/INS system by following an approach similar to that of Heng et al. [16, 17]. Here, the reference frame of the GNSS/INS system coincides with that of the IMU. We run semi-direct stereo visual odometry (VO) [15] for a stereo pair on each side of the vehicle. Each instance of stereo VO yields a set of camera poses and feature tracks. From hand-eye calibration using the reference camera’s poses from stereo VO and the GNSS/INS system’s poses, we obtain an initial estimate of the extrinsic transformation. Subsequently, we refine the extrinsic transformation by solving a non-linear least-squares problem in which we minimize the sum of squared reprojection errors associated with feature tracks while keeping the GNSS/INS system’s poses and inter-camera transformations fixed.

With the same fiducial target from intrinsic and extrinsic calibration of the multi-camera system, we perform extrinsic calibration between each LiDAR sensor and the calibrated multi-camera system. Given a set of images captured simultaneously from the multi-camera system, we detect the fiducial target in each image, and estimate its pose with respect to the multi-camera system by minimizing the squared sum of reprojection errors across all images in which the target was detected. At the same time, we identify the set of points corresponding to the fiducial target in the LiDAR scan by using plane segmentation [10], and estimate the target’s plane parameters with respect to the LiDAR sensor. With repeated observations of the target in different orientations, we independently estimate the rotation and translation components of the extrinsic transformation between the LiDAR sensor and the multi-camera system by doing singular value decomposition and solving a linear system of equations respectively. Subsequently, we refine the extrinsic transformation by minimizing the sum of squared point-plane errors; the points form part of the LiDAR scan identified as corresponding to the target, and the plane parameters are inferred from the estimated pose of the target with respect to the multi-camera system. Fig. 7 shows the projection of LiDAR scan points into a camera image using the results of extrinsic calibration between the LiDAR sensor and the multi-camera system.

IV. LOCALIZATION

One goal of Project AutoVision is to enable an autonomous vehicle to localize in both unmapped and premapped environments. A vehicle relying on map-based localization is restricted to movement within the map. We want to allow the vehicle to navigate beyond the map into unmapped areas by leveraging satellite imagery. However, a premapped environment enables the vehicle to localize with higher accuracy. Global pose estimates are susceptible to pose jumps; smooth local pose estimates are required for stable path tracking and to build consistent 3D maps. For this purpose, we use direct visual-inertial odometry which runs at the frame rate of the multi-camera system.

A. Direct Visual-Inertial Odometry

Our direct visual-inertial odometry (VIO) implementation for a multi-camera system [25, 26] estimates the local pose of the vehicle at 30 Hz. Our direct VIO implementation contains

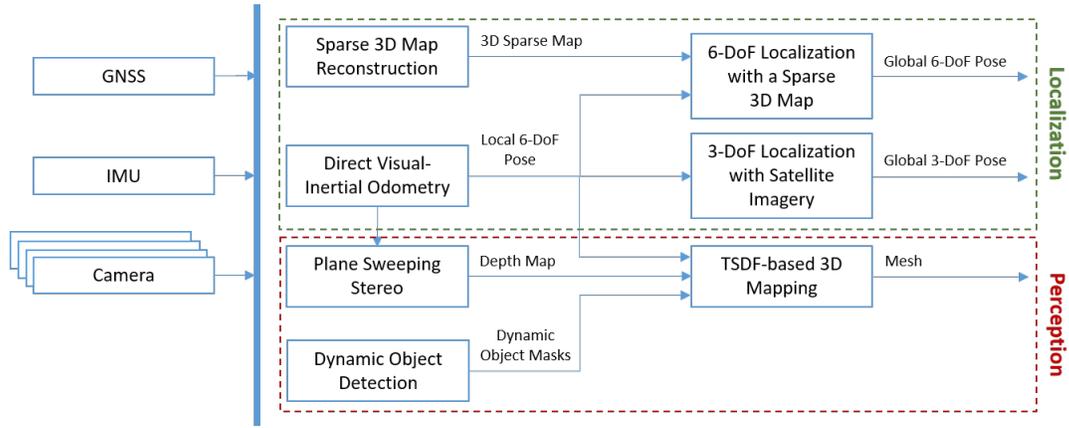


Fig. 5. Our AutoVision software architecture. GNSS is only used for sparse 3D map reconstruction.



Fig. 6. The fiducial target used for calibration of the multi-sensor system.

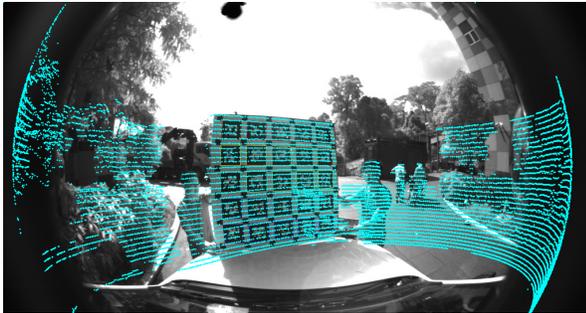


Fig. 7. Projection of LiDAR scan points into a camera image.

two threads: (1) the tracking thread estimates the local pose by minimizing photometric errors between the most recent keyframe and the current frame, and (2) the mapping thread initializes the depth of all sampled feature points using plane-sweeping stereo, and uses a sliding window optimizer to refine poses and structure jointly. Extensive experiments described by Liu et al. [26] show our implementation to work robustly for a 4-stereo-camera configuration with less than 1% translational drift in day-time and night-time with NIR illumination, and less than 2% translational drift in night-time without NIR illumination. Fig. 8 plots the pose estimates from our VIO implementation against ground truth data for a 8.2km route in a route covering both urban and rural environments.

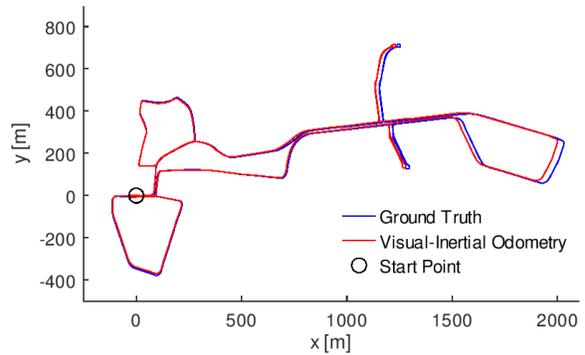


Fig. 8. The positions estimated by VIO vs ground truth positions. One front stereo pair and one rear stereo pair were used for VIO.

B. 3-DoF Localization with Satellite Imagery

In areas that have not been premaped, we rely on satellite imagery to estimate the 3-DoF global pose. Specifically, we estimate the (x, y) position and heading of the vehicle with respect to the UTM coordinate frame. Assuming that the initial position and heading are known from user input, we use the particle filter approach in which we use local pose data from VIO for particle propagation and output from a deep network [19, 18] for particle weighting.

As shown in Fig. 9, the deep network called CVM-Net is a Siamese network that takes satellite and ground-level panoramic images as input. We obtain the panoramic image by stitching the cylindrical projections of images taken from four cameras: one camera on each side of the vehicle. For each image, we extract local features via fully convolutional networks. Two aligned NetVLADs aggregate local features from both images into global descriptors that are in a common space for similarity comparison. The weight for each particle is inversely proportional to the Euclidean distance between the global descriptors corresponding to the ground-level panoramic image and the satellite image patch nearest the particle's position.

We run two experiments with a 5km route in both an urban environment and a rural environment. Experimental

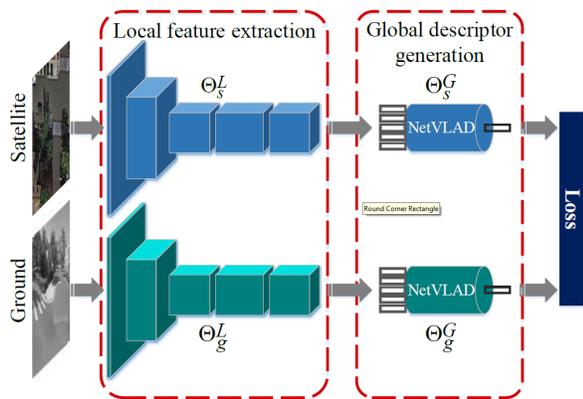


Fig. 9. The architecture of our deep network, CVM-Net, for cross-view matching [19, 18].

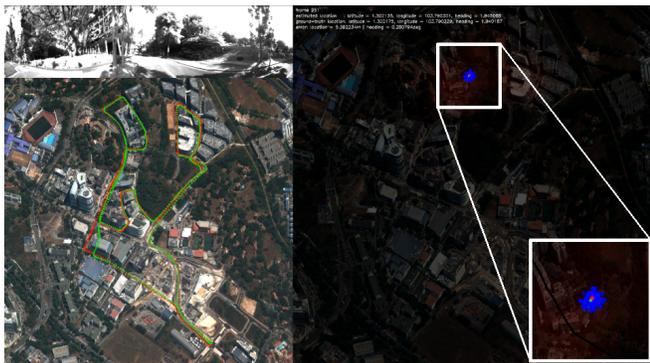


Fig. 10. A visualization of our satellite-imagery-based localisation in an urban environment.

results show that our satellite-imagery-based localization achieves an average position error of 9.92m and an average heading error of 0.32° over a 5km route in the urban environment, and an average position error of 9.29m and an average heading error of 0.42° in the rural environment. Fig. 10 visualizes our satellite-imagery-based localisation in an urban environment. The top left image shows the ground-view panoramic image. The bottom left image shows the paths estimated by our localization and GNSS/INS system in green and red respectively on the bottom left. In the right image, particles are shown in blue on the right and interposed against a likelihood map; the more red the pixel, the higher the likelihood that the vehicle is located at that pixel.

C. Sparse 3D Map Reconstruction

Sparse 3D map reconstruction is required for map-based 6-DoF localization which is described in Section IV-D. Prior to localization, we build a sparse 3D map in which each 3D point is associated with one or more local SIFT features [27].

To minimize the time required for large-scale reconstruction, our approach does not reconstruct the scene from scratch, and instead, uses reasonably accurate initial pose estimates from a GNSS/INS system to initialize all camera poses. In addition, we require a minimum amount of camera motion between images used for mapping. Next, we perform feature matching between nearby images, and use the feature

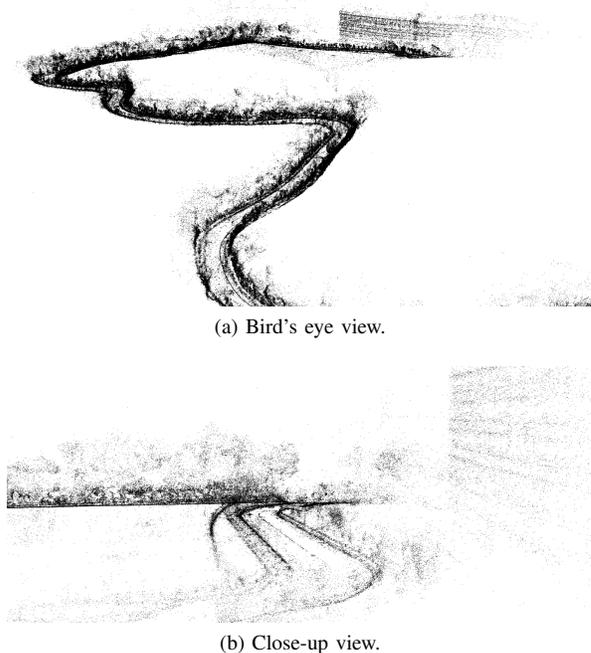


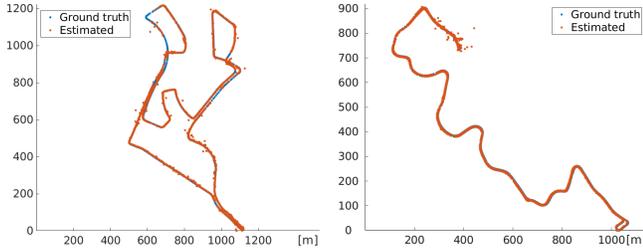
Fig. 11. The sparse 3D map of a mixed urban and rural environment and generated by our reconstruction pipeline.

matches and initial poses to triangulate the scene. We then repeatedly optimize the scene structure and camera poses using bundle adjustment followed by the merging of feature tracks. This approach is implemented on top of the COLMAP structure-from-motion (SfM) framework [30]. During bundle adjustment, we enforce that the extrinsic parameters of the multi-camera system on the AutoVision vehicle remain constant. Fig. 11 shows a sparse 3D map of a mixed urban and rural environment.

D. 6-DoF Localization with a Sparse 3D Map

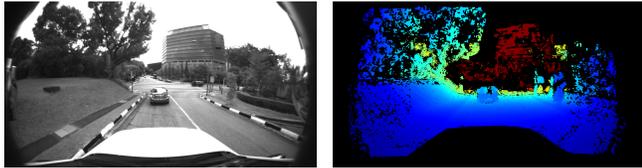
Given a sparse 3D map computed using the approach described in Section IV-C and without prior knowledge of the AutoVision vehicle's global pose, we localize the AutoVision vehicle by extracting local SIFT features [27] from the images captured by the cameras mounted on the AutoVision vehicle and matching the descriptors of these 2D features against the descriptors associated with the 3D points in the map. From the 2D-3D matches, we apply a generalized perspective-n-point pose solver [22, 23] inside a RANSAC loop [8] to estimate the vehicle's pose. With a sparse 3D map of a large area which contains many 3D points, 2D-3D matching is the main computational bottleneck in our pipeline. In this case, we use a prioritized matching approach based on Active Search [29] for improved matching efficiency. Geppert et al. [11] describe our localization approach in greater detail.

For our experiments, we use the same routes used for the satellite-imagery-based localization experiments described in Section IV-B. Our localization pipeline runs at around 2 Hz on the AutoVision vehicle. Fig. 12 shows the estimated and ground truth positions for the urban and rural routes. For the urban route, the mean and median errors of all reported

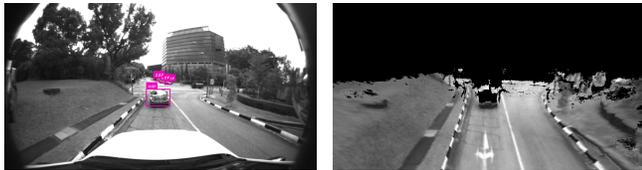


(a) Urban environment. In some parts of the area, the localization fails consistently to estimate a pose. (b) Rural environment. The errors at the top end are likely caused by repetitive structures on both sides of the road.

Fig. 12. The positions estimated by map-based localization vs. ground truth positions for a 5km route in two different environments.



(a) Image from the front center camera. (b) Depth image generated by plane-sweeping stereo.



(c) Dynamic objects detected by the object detector. (d) Raycasted TSDF volume.

Fig. 13. Output from the AutoVision perception modules [5].

poses are 3.31m and 1.84m for the position, and 2.6° and 1.9° for the heading, respectively. For the rural route, the mean and median errors of all reported poses are 3.48m and 1.81m for the position, and 4.2° and 3.3° for the heading, respectively.

V. PERCEPTION

Dense 3D mapping in real-time is a prerequisite for autonomous navigation in complex environments. Our perception pipeline begins with plane-sweeping stereo generating depth images. In turn, depth images are fused into a truncated signed distance function (TSDF) volume. A 3D map is reconstructed from this TSDF volume via ray-casting. To avoid dynamic objects from corrupting the 3D map via trails of artefacts, we detect potentially dynamic objects not belonging to the static environment, and remove their associated depth estimates from the depth images prior to depth fusion. Fig. 13 visualizes the outputs from plane-sweeping stereo, object detection, and TSDF-based 3D mapping. More details of our perception pipeline can be found in [5].

A. Plane-Sweeping Stereo

Plane-sweeping stereo computes a depth image for multiple images with known camera poses by sweeping a set

of planes through 3D space. Each plane represents a depth hypothesis and defines a homography mapping from every other view to the reference view. We estimate the depth for each pixel in a reference image by using each plane to warp each non-reference image to the reference image, evaluating the image dissimilarity at that pixel, and choosing the plane that minimizes the image dissimilarity. We use the GPU implementation of plane-sweeping stereo for fisheye cameras [13] which computes depth images directly from fisheye images without the need for undistortion, allowing us to use the full field-of-view of the cameras. On the AutoVision vehicle, plane-sweeping stereo runs at an average of 15 Hz for the 5 cameras at the front of the vehicle and with images downsampled to half-resolution.

B. TSDF-based 3D Mapping

A single depth image may not contain sufficient geometric information for up-stream modules such as 3D semantic segmentation and motion planning. Hence, we need to fuse depth images estimated at different positions in time to create a dense 3D map. We use a standard fusion technique: the scene is represented via a set of voxels where each voxel stores a TSDF value [6]. Here, each voxel stores the signed distance to the closest object surface (negative inside of objects, positive outside of objects, zero on surfaces), truncated to a certain maximum / minimum value. Whenever a new depth image along with its camera pose becomes available, we update the 3D model. We use the map fusion pipeline in the InfiniTAM library [20, 21]. We also use the fast raycasting algorithm in [20, 21] to reconstruct the 3D map in the current camera view. The pipeline runs at around 20 Hz on the AutoVision vehicle.

C. Dynamic Object Detection

Dynamic objects leave behind trails of artefacts in the 3D map. We leverage 2D object detection to solve this problem. Given a reference image, we detect dynamic objects, i.e. humans and vehicles. In turn, for the corresponding depth image, we mask out pixels located within the 2D bounding boxes of detected objects. This way, we avoid integrating depth estimates associated with dynamic objects into the 3D map. We use the YOLOv3 object detection network [28] trained on the Microsoft COCO dataset [24]. To improve the inference performance with distorted grayscale images from our NIR fisheye cameras, we fine-tune the network by truncating the first and last layers, and retrain the network using our labeled datasets.

VI. CONCLUSIONS

Project AutoVision has successfully demonstrated localization and 3D scene perception for autonomous vehicles with multi-camera systems, in both urban and rural environments, and without GNSS. As Project AutoVision progresses, we will continue to enhance localization and perception capabilities, and add more modules to our software stack. These modules include but are not limited to, change detection for 3D maps, obstacle detection, dynamic object tracking and classification, and semantic 3D mapping.

REFERENCES

- [1] J. Baretto and H. Araujo. Issues on the geometry of central catadioptric image formation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2001.
- [2] M. Bojarski, D. D. Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang, X. Zhang, J. Zhao, and K. Zieba. End to end learning for self-driving cars. *CoRR*, abs/1604.07316, 2016.
- [3] A. Broggi, P. Cerri, S. Debattisti, M. C. Laghi, P. Medici, D. Molinari, M. Panciroli, and A. Prioletti. Proudpublic road urban driverless-car test. *IEEE Transactions on Intelligent Transportation Systems*, 16(6): 3508–3519, 2015.
- [4] D. C. Brown. Decentering distortion of lenses. *Photometric Engineering*, 32(3):444–462, 1966.
- [5] Z. Cui, L. Heng, Y. C. Yeo, A. Geiger, M. Pollefeys, and T. Sattler. Real-time dense mapping for self-driving vehicles using fisheye cameras. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2019.
- [6] B. Curless and M. Levoy. A volumetric method for building complex models from range images. In *Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, 1996.
- [7] J. Engel, V. C. Usenko, and D. Cremers. A photometrically calibrated benchmark for monocular visual odometry. *CoRR*, abs/1607.02555, 2016.
- [8] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24:381395, 1981.
- [9] P. Furgale, U. Schwesinger, M. Rufli, W. Derendarz, H. Grimmert, P. Mühlfellner, S. Wonneberger, J. Timmer, S. Rottmann, B. Li, B. Schmidt, T. N. Nguyen, E. Cardarelli, S. Cattani, S. Brüning, S. Horstmann, M. Stellmacher, H. Mielenz, K. Köser, M. Beermann, C. Häne, L. Heng, G. H. Lee, F. Fraundorfer, R. Iser, R. Triebel, I. Posner, P. Newman, L. Wolf, M. Pollefeys, S. Brosig, J. Effertz, C. Pradalier, and R. Siegwart. Toward automated driving in cities using close-to-market sensors: An overview of the v-charge project. In *IEEE Intelligent Vehicles Symposium (IV)*, 2013.
- [10] A. Geiger, F. Moosmann, U. Car, and B. Schuster. Automatic camera and range sensor calibration using a single shot. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2012.
- [11] M. Geppert, P. Liu, Z. Cui, M. Pollefeys, and T. Sattler. Efficient 2d-3d matching for multi-camera visual localization. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2019.
- [12] C. Geyer and K. Daniilidis. A unifying theory for central panoramic systems and practical implications. In *European Conference on Computer Vision (ECCV)*, 2000.
- [13] C. Häne, L. Heng, G. H. Lee, A. Sizov, and M. Pollefeys. Real-time direct dense matching on fisheye images using plane-sweeping stereo. In *International Conference on 3D Vision (3DV)*, 2015.
- [14] C. Häne, L. Heng, G. H. Lee, F. Fraundorfer, P. Furgale, T. Sattler, and M. Pollefeys. 3d visual perception for self-driving cars using a multi-camera system: Calibration, mapping, localization, and obstacle detection. *Image and Vision Computing (IVC)*, 68:14–27, 2017.
- [15] L. Heng and B. Choi. Semi-direct visual odometry for a fisheye-stereo camera. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2016.
- [16] L. Heng, P. Furgale, and M. Pollefeys. Leveraging image-based localization for infrastructure-based calibration of a multi-camera rig. *Journal of Field Robotics (JFR)*, 32:775–802, 2015.
- [17] L. Heng, G. H. Lee, and M. Pollefeys. Self-calibration and visual slam with a multi-camera system on a micro aerial vehicle. *Autonomous Robots (AURO)*, 39:259–277, 2015.
- [18] S. Hu and G. H. Lee. Image-based geo-localization using satellite imagery. *International Journal of Computer Vision (IJCV)*, 2019.
- [19] S. Hu, M. Feng, R. M. H. Nguyen, and G. H. Lee. Cvmnet: Cross-view matching network for image-based ground-to-aerial geo-localization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [20] O. Kähler, V. A. Prisacariu, C. Y. Ren, X. Sun, P. Torr, and D. Murray. Very high frame rate volumetric integration of depth images on mobile devices. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 21(11):1241–1250, 2015.
- [21] O. Kähler, V. Prisacariu, J. Valentin, and D. Murray. Hierarchical voxel block hashing for efficient integration of depth images. *IEEE Robotics and Automation Letters (RA-L)*, 1(1):192–197, 2016.
- [22] L. Kneip, P. Furgale, and R. Siegwart. Using multi-camera systems in robotics: Efficient solutions to the npnp problem. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2013.
- [23] G. H. Lee, B. Li, M. Pollefeys, and F. Fraundorfer. Minimal solutions for the multi-camera pose estimation problem. *International Journal of Robotics Research (IJRR)*, 34:837–848, 2015.
- [24] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, 2014.
- [25] P. Liu, L. Heng, T. Sattler, A. Geiger, and M. Pollefeys. Direct visual odometry for a fisheye-stereo camera. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2017.
- [26] P. Liu, M. Geppert, L. Heng, T. Sattler, A. Geiger, and M. Pollefeys. Towards robust visual odometry with a multi-camera system. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*,

2018.

- [27] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision (IJCV)*, 60(2):91110, 2004.
- [28] J. Redmon and A. Farhadi. Yolov3: An incremental improvement. *CoRR*, abs/1804.02767, 2018.
- [29] T. Sattler, B. Leibe, and L. Kobbelt. Efficient & effective prioritized matching for large-scale image-based localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 39:17441756, 2017.
- [30] J. Schönberger and J. Frahm. Structure-from-motion revisited. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [31] U. Schwesinger, M. Bürki, J. Timpner, S. Rottmann, L. Wolf, L. M. Paz, H. Grimmer, I. Posner, P. Newman, C. Häne, L. Heng, G. H. Lee, T. Sattler, M. Pollefeys, M. Allodi, F. Valenti, K. Mimura, B. Goebelsmann, W. Derendarz, P. Mühlfellner, S. Wonneberger, R. Waldmann, S. Grysczyk, C. Last, S. Brüning, S. Horstmann, M. Bartholomäus, C. Brummer, M. Stellmacher, F. Pucks, M. Nicklas, and R. Siegwart. Automated valet parking and charging for e-mobility. In *IEEE Intelligent Vehicles Symposium (IV)*, 2016.
- [32] J. Wang and E. Olson. AprilTag 2: Efficient and robust fiducial detection. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2016.
- [33] J. Ziegler, P. Bender, M. Schreiber, H. Lategahn, T. Strauss, C. Stiller, T. Dang, U. Franke, N. Appenrodt, C. G. Keller, E. Kaus, R. G. Herrtwich, C. Rabe, D. Pfeiffer, F. Lindner, F. Stein, F. Erbs, M. Enzweiler, C. Knoppel, J. Hipp, M. Haueis, M. Trepte, C. Brenk, A. Tamke, M. Ghanaat, M. Braun, A. Joos, H. Fritz, H. Mock, M. Hein, and E. Zeeb. Making bertha drive - an autonomous journey on a historic route. *IEEE Intelligent Transportation Systems Magazine*, 6(2):8–20, 2014.