An Improved Approach to Omnidirectional Stereo Vision for Environment Mapping in Future Driverless Vehicles

Michael Groom supervised by Prof. T.P. Breckon

***Abstract*—Several environmental mapping technologies are used in automated driving systems, including LiDAR, radar and cameras. In particular, cameras used in stereo vision systems provide a good compromise between accuracy and cost. Most recent stereo vision studies use conventional cameras with a limited field of view. Stereo vision systems using omnidirectional cameras have been proposed to overcome this problem. However, few attempts have been made to investigate omnidirectional stereo depth accuracy.**

This paper presents an omnidirectional stereo system producing 360° disparity maps at up to 14.4 FPS with improved disparity coverage than previous work. The accuracy of using a spherical camera model combined with a longitude-latitude projection for omnidirectional stereo is investigated, showing error in calculated depth increases significantly as the angle from the cameras optical axis approaches the limit of the cameras field of vision.

A separate approach is outlined in this paper using a perspective undistortion function and the conventional pinhole camera model, allowing omnidirectional cameras to be used in conventional stereo systems. The proposed approach exhibits improved depth accuracy at large angles from the cameras optical axis compared to omnidirectional stereo systems that use a spherical camera model.

I. Introduction

AUTOMATED Driving Systems (ADS) have the potential to increase the safety of road users massively. The NHTSA reported that in the United States in 2018, over 36,500 people died in motor vehicle-related crashes, with a staggering 94% of these accidents caused by human error [1]. Vehicles with ADS could reduce the number of fatalities by addressing the root cause of most crashes, the human driver. The same technology could also help to improve traffic flow [2], reduce carbon emissions if integrated into shared autonomous vehicles [3] and increase the mobility and personal freedoms of those who cannot drive.

Some form of environmental mapping technology is required to create a vehicle with ADS. For ADS technology to become widespread, it must be affordable. Therefore the environmental mapping technology used must be cost-effective. Different environmental mapping technologies are currently used for self-driving applications, including LiDAR, radar, ultrasonic sensors and stereo/multicam systems, herein referred to as cameras. Each of these technologies come with advantages and disadvantages. For example, LiDAR, which is used by companies such as Daimler [4] and Google [5] as their main mapping technology, is prohibitively expensive for most vehicle owners with sensors costing around $75,000 [6]. In this study, omnidirectional cameras are used to create a cost-effective stereo vision system, and its accuracy is investigated.

The prior work of Lin and Breckon [7] made use of the spherical projection model of Mei and Rives [8] to produce disparity maps from omnidirectional images. Notably, no attempt to evaluate depth calculated from disparity maps produced was shown.

Contributions made in this paper address problems identified in previous work by Lin [9]. Firstly, a new camera configuration is developed with greater stability to reduce noise introduced into the system from vibrations from the road surface. The new configuration has improved accessibility to the cameras and keeps blind spots in areas of no interest to automotive depth sensing. Secondly, the spherical camera model proposed by [8], used in [7] for omnidirectional stereo vision, is implemented, and its accuracy for use in a stereo system is investigated, revealing it suffers from significant depth error at wide angles from the optical axis. This paper adopts the perspective undistortion function proposed by [10] to overcome this problem, producing an omnidirectional stereo vision system with improved depth accuracy than [7] at large angles from than optical axis. Finally, using two off-the-shelf cameras, real-time omnidirectional stereo is achieved using both proposed camera models, producing disparity maps at up to 14.4 FPS (frames per second) with improved disparity coverage than [7].

II. Related Work

In this section, common environmental mapping technologies used for ADS are compared to justify investigating an omnidirectional stereo vision system for automotive depth sensing. Next, different camera models are discussed that allow for omnidirectional cameras to be utilised in a stereo vision system. This is followed by a brief discussion on stereo correspondence algorithms. Finally, existing omnidirectional stereo vision systems are examined.

A. Environmental Mapping

A comparison between the most common forms of environmental mapping used in driver-less vehicles is presented in Table I. If ADS is to become commonplace on the road, the mapping technology it uses must be passive to avoid interference from other road users. Table I shows that cameras are the only viable low-cost option for long-range environmental mapping. For this reason, future ADS systems are likely to rely solely on cameras for depth sensing. This prediction is supported by recent news that the Tesla Full Self-Driving Beta will soon be transitioning to a complete camera-based

	LiDAR [6]	Radar [11]	Cameras [12]	Ultrasonic [13]
Active / Passive	Active	Active	Passive	Active
Range (m)	120	250	500	5
Horizontal FOV (°)	360	20	120	120
Vertical FOV (°)	26.9	4.5	120	120
Angular Resolution (°)	0.4	0.1	-	3
Linear Resolution (cm)	-	10	-	2
Accuracy	2 cm	10 cm	-	10%
Cost (USD)	75,000	3,500	450	670

TABLE I. Comparison of popular environmental mapping technologies. Gaps represent values that are product dependant.

approach [14]. Conventional cameras have limited a field of vision (FOV); therefore, several cameras are required to create a complete 360° FOV, leading to an increased overall cost. By considering omnidirectional cameras, which typically have a FOV greater than 180°, the cost of creating a system with a complete 360° FOV could be reduced as fewer cameras would be required.

B. Camera Models

Camera models that allow for omnidirectional cameras to be used in a stereo vision system must first be considered before implementing such a system. Firstly, a definition of what classifies an omnidirectional camera is given by Puig et al. [15], who separate omnidirectional cameras into two categories: central and non-central. Examples of non-central cameras include rotating cameras, which rotate along a circular trajectory taking several images to generate a complete 360° FOV. Other examples of non-central cameras include poly-cameras, which are clusters of conventional cameras pointed in different directions, and dioptric systems that use a wide-angle lens like a fish-eye lens combined with a conventional camera. The proposed omnidirectional stereo rig uses two Ricoh Theta S cameras [16], which use a dual fish-eye lens. Therefore camera models that can be applied to cameras that use fish-eye lens are considered.

Scaramuzza et al. [10] present a camera calibration technique for omnidirectional cameras, assuming that omnidirectional images are distorted images and propose an image projection function to undistort them. The image projection function is a Taylor series expansion, whose coefficients are found during a calibration process, similar to the seminal work of Zhang [17]. This calibration process was improved upon by Urban et al. [18]. This work allows spherical images produced by omnidirectional cameras to be treated as planar images.

Geyer and Daniilidis [19] provide a basic projection model for panoramic images. Barreto and Araujo [20] establish a general model for image formation in central catadioptric images. Mei and Rives [8] took these models and formed a unified projection model. A new parameter, ξ, is proposed that compensates for any difference between the idealised spherical camera model centre and the actual camera centre. This parameter is found during a calibration process similar to work by Zhang [17]. Li et al. [21] proposed a similar model to Mei and Rives [8]. However, the model proposed by Li et al. [21] did not include a term to model differences in ideal and real camera centres similar to the ξ parameter outlined by Mei and Rives [8]. Li et al. [21] used their proposed spherical camera model to reformulate the conventional planar stereo problem for spherical cameras by defining disparity and depth for a spherical stereo system.

C. Stereo Matching Algorithms

For stereo vision to be applied for automotive depth sensing, a stereo correspondence algorithm must be chosen that is accurate and can also run in real-time. There is a trade-off between speed and accuracy when choosing a stereo correspondence algorithm for automotive sensing. The KITTI stereo 2015 benchmark [22] serves as a leaderboard for the current state of the art in stereo correspondence algorithms. It pushes the need for algorithms to perform well in automotive sensing applications by testing algorithms on dynamic outdoor scenes. PSMNet [23] was chosen due to its good trade-off between accuracy and speed when evaluated on the KITTI stereo 2015 leaderboard [22] and its open-source implementation. Semi-Global Block Matching (SGBM) [24] was also selected as it was chosen for the omnidirectional stereo system produced by Lin and Breckon [7], and allowed the accuracy of such a system to be evaluated. Finally, SGBM was also implemented with a weighted least squares filter to improve upon disparity maps produced by SGBM, which are prone to noise.

D. Omnidirectional Stereo

Goa and Shen [25] use two large FOV fish-eye lenses facing opposite directions to create full spherical monocular coverage. Overlapping regions are rectified into stereo image pairs and used for depth estimation. A lens-specific calibration procedure is used for this approach, which means these results are not easily transferable.

Won et al. [26] propose a neural network model for omnidirectional depth estimation from a multi-view stereo setup. Their model extracts features from four omnidirectional cameras and uses spherical sweeping and cost volume aggregation to produce omnidirectional disparity maps. Since this method differs significantly from the conventional approach of using rectified images for conventional stereo algorithms by using spherical sweeping, it is not used in this study.

Ma et al. [27] use the spherical camera model proposed by [21] to create a spherical stereo system. The disparity maps produced by Ma et al. [27] *"look a little messy"*. This work was improved upon by Lin and Breckon [7], who used the spherical camera model of Mei and Rives [8]. However, no attempt to verify the accuracy of disparity maps was made using this method. In this work, the accuracy of a spherical stereo system using the same technique as that used in [7], as well as the accuracy of a spherical stereo system using the undistortion technique proposed by Scaramuzza et al. [10] are investigated.

III. Theory

In this section, pinhole and spherical camera models are introduced. Subsequently, conventional and omnidirectional

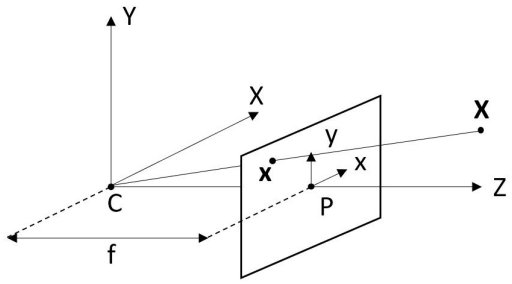


Fig. 1. Pinhole camera model. This camera model is used in conventional stereo vision systems.

stereo system theory is presented, followed by two separate projection techniques for omnidirectional cameras. Finally, epipolar geometry and calculating depth from disparity using conventional and spherical camera models are summarised.

A. Pinhole and Spherical Camera Models

The pinhole camera model is typically used to model how a conventional camera depicts a 3D scene. The pinhole camera model can be used to project a point $\mathbf{X} = (X, Y, Z)^T$ in the world coordinate system to a point $\mathbf{x}$ on a plane, as shown in Figure 1. This plane is known as the image plane. The centre of projection is known as the optical centre, C, and is assumed to be at the origin of the Euclidean coordinate space. The line from the optical centre that meets perpendicular to the image plane is known as the optical axis, and where the optical axis meets the image plane forms the principal point, P. The distance between the optical centre C and the principal point P is known as the focal length f. Using homogeneous coordinates, the point $\mathbf{X}$ is mapped to the point $\mathbf{x}$ in the image plane by:

$$\mathbf{x} = [K \mid 0]\mathbf{X}\,, \tag{1}$$

where K is the camera matrix. K is defined as:

$$K = \begin{bmatrix} f & 0 & u_0 \\ 0 & f & v_0 \\ 0 & 0 & 1 \end{bmatrix}\,, \tag{2}$$

where (u_0, v_0) are the coordinates of the principal point P. The spherical camera model proposed by Mei and Rives [8] can be used to represent an omnidirectional camera. It differs from a conventional pinhole camera model as instead of world points being projected onto a planar surface, points are projected onto a spherical surface, centred around the camera centre. Firstly, a point $\mathbf{X} = (X, Y, Z)^T$ is projected onto a unit sphere by

$$X_s = \frac{X}{\|X\|}\,. \tag{3}$$

The point is then changed to a new reference frame centred at $Cp = (0, 0, \xi)^T$. The point $\mathbf{X} = (X, Y, Z)^T$ becomes $\mathbf{X} = (X, Y, Z{+}\xi)^T$ in the new reference frame. The parameter ξ is used to model the difference between the spherical model and real camera centres. This point is then projected to the point m on the normalised plane, as shown in Figure 2, with coordinates $m = (\frac{X}{Z+\xi}, \frac{Y}{Z+\xi}, 1)^T$. Finally, a generalised camera matrix K is used to project the point m from the normalised plane to the image plane. K is defined as:

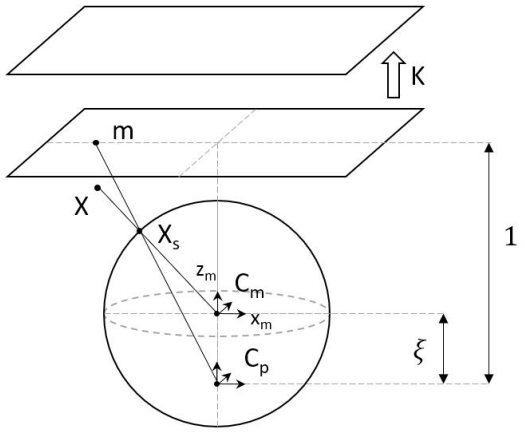


Fig. 2. Spherical camera model proposed by Mei and Rives [8].

$$K = \begin{bmatrix} f_1\eta & f_1\eta\alpha & u_0 \\ 0 & f_2\eta & v_0 \\ 0 & 0 & 1 \end{bmatrix}\,, \tag{4}$$

where again (u_0, v_0) are the coordinates of the principal point P, f_1 and f_2 are the focal lengths in pixels, α is the skew and η is dependant on lens geometry [8].

In reality, a camera uses a lens to focus light rays, which introduces radial and tangential distortions [28]. The radial distortion can be approximated by:

$$\begin{aligned} u_i &= u_d \left(1 + k_1 r^2 + k_2 r^4\right)\,, \\ v_i &= v_d \left(1 + k_1 r^2 + k_2 r^4\right)\,, \end{aligned} \tag{5}$$

where $r = \sqrt{u_d^2 + v_d^2}$. Tangential distortion can be modelled by:

$$\begin{aligned} u_i &= u_d + \left[2k_3 u_d v_d + k_4 \left(r^2 + 2u_d^2\right)\right]\,, \\ v_i &= v_d + \left[k_3 \left(r^2 + 2v_d^2\right) + 2k_4 u_d v_d\right]\,. \end{aligned} \tag{6}$$

The distortion coefficients $D = (k_1, k_2, k_3, k_4)$ along with the camera matrix K are known as the camera intrinsics, which are unique to a single camera and are found during camera calibration.

B. Stereo Vision

In a conventional stereo system, two cameras are used to observe the same scene. By matching corresponding pixels in each image, depth information can be calculated. Any pixel in one image could correspond to any pixel in the other, which means finding matches is computationally expensive and prone to mismatching for large images. Epipolar geometry is used to reduce the search space, which constrains corresponding pixels to the same image row, and is discussed later in Section III-C. This is expressed mathematically as:

$$\{x_1\}^T [\mathbf{F}] \{x_2\} = 0\,, \tag{7}$$

where x_1 and x_2 are the same world point viewed from image planes 1 and 2, and $\mathbf{F}$ is the fundamental matrix. For a stereo system using conventional cameras, the disparity is defined as the difference in location between two corresponding points along an image row:

$$d = u_l - u_r\,, \tag{8}$$

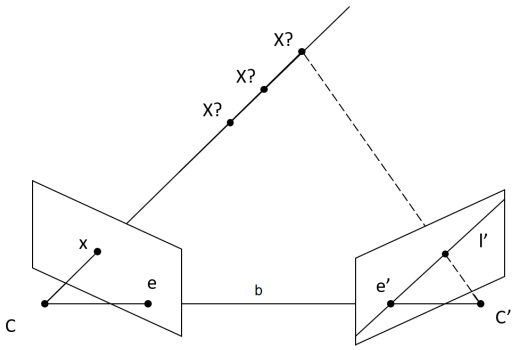


Fig. 3. Epipolar geometry of a conventional planar stereo system.

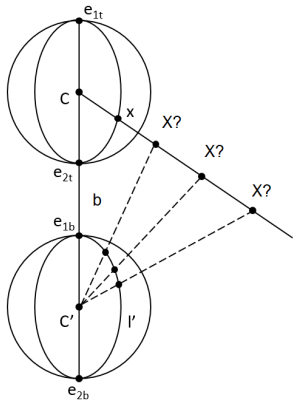


Fig. 4. Spherical epipolar geometry of a vertical binocular spherical stereo system.

where u_l and u_r are corresponding pixels in the left and right image planes after image rectification. For a spherical stereo system, the disparity can no longer be expressed as a linear distance and instead represents a change in arc length:

$$d = u_l - u_r = f_s(\phi_l - \phi_r)\,, \tag{9}$$

where f_s is the radius of the spherical camera model.

C. Epipolar Geometry

Epipolar geometry constrains corresponding pixels to the same image rows in rectified images, significantly reducing the computational cost of producing dense disparity maps. Epipolar geometry involving planar image planes is shown in Figure 3 and spherical image planes in Figure 4. Assuming there is a point in space $\mathbf{X} = (X, Y, Z)^T$ that is visible from both cameras, there must exist vectors from $\mathbf{X}$ to the centre of each camera, C and C' respectively. The camera baseline, b, is the vector between the two camera centres C and C'. These three vectors form a plane known as the epipolar plane. The epipolar plane intersects the two image planes in lines known as epipolar lines. As seen in Figure 3, epipolar lines are straight lines when intersecting with planar image planes. Epipolar lines appear as conics when intersecting with spherical image planes, as shown in Figure 4. The fundamental matrix $\mathbf{F}$ is the algebraic representation of epipolar geometry, as shown in Eq. (7), and is used to map points between two image planes. The fundamental matrix encapsulates the stereo systems intrinsic and extrinsic parameters and is found during camera calibration. The extrinsic parameters are composed of a rotation and translation and describe the relationship between the two image planes. Extrinsic parameters can be used to rectify images, which involves transforming images so that the two cameras optical axes are perfectly aligned. Once images are rectified, epipolar lines appear as image rows, greatly simplifying the task of finding correspondences between the two images.

D. Perspective Undistortion of Spherical Images

Work by [10] allows omnidirectional cameras to be treated as conventional cameras with planar image planes by proposing that a Taylor series can be used to approximate a mapping function to project spherical images to planar images. A 2D image point $\mathbf{m} = [u, v]^T$ can be mapped to its corresponding scene point $\mathbf{X_c} = [X_c, Y_c, Z_c]^T$ in the camera coordinate system through the imaging function g:

$$\mathbf{X_c} = \lambda g(\mathbf{m}) = \lambda(u, v, f(\rho))^T\,, \tag{10}$$

with $\lambda > 0$ and $\rho = \sqrt{u^2 + v^2}$ is the radial Euclidean distance to from the image centre. Instead of defining a specific mapping function g, [10] instead approximated it with a Taylor series:

$$f(\rho) = a_0 + a_2\rho^2 + \cdots + a_n\rho^n\,, \tag{11}$$

where coefficients $a_0, a_2, \ldots, a_n$ are found during camera calibration. This imaging function allows conventional stereo techniques to be applied to omnidirectional images. However, to represent a hemispherical view using a perspective projection an infinitely large image is required, which results in this imagining function limiting the FOV of the omnidirectional camera to less than $180°$.

E. Spherical Projection

Most recent stereo matching algorithms use rectangular rectified images and therefore search for correspondences along corresponding image rows. This is because epipolar lines appear as image rows in rectified images, reducing the correspondence problem from 2D to 1D. However, in spherical images, epipolar lines appear as conics around the image centre, meaning that most stereo matching algorithms cannot be applied to systems using spherical camera models. To use these algorithms, images must first be transformed so that epipolar lines no longer appear as conics but as straight lines. Li et al. [21] proposed that a longitude-latitude projection could be used to transform spherical images, making conic epipolar lines appear as straight lines, allowing conventional stereo correspondence algorithms to be applied to omnidirectional stereo. A longitude-latitude projection is defined as:

$$\begin{aligned} u &= f_s\theta\,, \\ v &= f_s\phi\,, \end{aligned} \tag{12}$$

where θ and ϕ are the polar and azimuth angles if the two epipoles in the spherical image are defined as two poles of the coordinate system. As noted by Li et al. [21] when first proposing the use of longitude-latitude projections for transforming spherical images, this approach produces regions near the epipoles where the error in estimated depth becomes large. The epipole, e, is the point at which the baseline

intersects the spherical image plane and is shown in Figure 4.

F. Depth from Disparity

When using conventional cameras calculating depth from disparity is relatively simple, with the relation given by:

$$z = \frac{bf}{d}\,, \tag{13}$$

where z is the depth, b is the baseline distance between the two cameras, f is the focal length of the cameras, and d is the disparity. Calculating depth from spherical disparity is more complex and is briefly presented below. A full derivation for depth from a vertical spherical stereo system is available in [7]. The distance from a point to the top spherical camera centre is given by:

$$\rho_t = b\frac{\sin\left(v_b/f_s\right)}{\sin\left(d/f_s\right)}\,, \tag{14}$$

where v_b is the vertical pixel coordinate in the bottom image, d is the vertical disparity.

IV. Methods

A new camera configuration was implemented for this study and is discussed in this section. This is followed by two different methods of camera calibration. Subsequently, the stereo matching algorithms used are discussed. Finally, an experiment is proposed to evaluate the accuracy of the different proposed camera calibration methods.

A. Camera Configuration

In previous work from Lin [9], the system configuration suffered from noise generated by cameras being shaken by vibrations from the vehicle and road surfaces. A new system configuration is proposed to address this. The new configuration places the cameras in a similar top-bottom configuration to Lin [9]. However, the two cameras are now also horizontally offset from each other, as seen in Figure 5, for greater accessibility to the cameras while keeping blind spots in areas of no interest to automotive depth sensing. Due to this offset, source images are rotated before being rectified to make this configuration a vertical stereo system. This configuration makes use of a perspex hemisphere instead of the cylinder used in [9], which allows for extra room to attach a larger baseplate and hence allowed for more mounts to attach to the vehicle over a larger area for better stability. The configuration makes use of two Ricoh Theta S cameras in their live mode [16] and are interfaced to a computer using OpenCV [29] via USB 3.0 cables. Each Ricoh Theta S camera consists of two fish-eye cameras, with each fish-eye camera having a FOV of approximately $190°$, combining to create a complete $360°$ FOV. The same notation as [9] will be used, with cameras being denoted as Top-Front (tf), Top-Back (tb), Bottom-Front (bf), and Bottom-Back (bb). Dual fish-eye images from the two cameras are split to create four 640×640 images. The cameras are connected via USB 3.0 to a laptop computer (Intel Core i7-6700HQ CPU at 2.60GHz, GeForce GTX 970M, CUDA 5.2) running Windows 10.

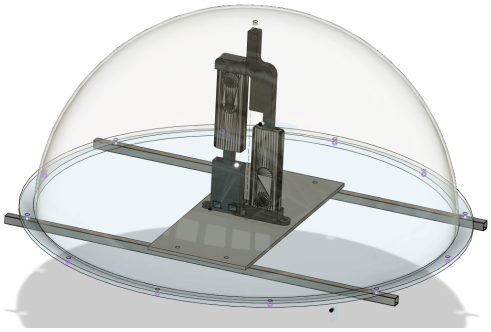


Fig. 5. 3D CAD model of the omnidirectional camera rig.

B. Camera Calibration

Two separate calibration techniques were implemented for this study. The first approach used the spherical camera model proposed by Mei and Rives [8], herein referred to as Mei's Calibration Method. The second used a conventional planar stereo system with the imaging function proposed by Scaramuzza et al. [10], herein referred to as Scaramuzza's Calibration Method.

Coefficients for the Taylor series that describes the imaging function in Eq. (11) were found using the OCamCalib MATLAB toolbox [30] with the ImprovedOCamCalib extension [18].

Next, stereo calibration was performed for both calibration methods. Separate stereo calibrations were performed for each scale factor investigated when using Scaramuzza's Calibration Method. To estimate intrinsic and extrinsic camera parameters for both calibration techniques a Levenberg-Marquardt Algorithm (LMA) optimisation technique was used. The LMA algorithm is an iterative approach where reprojection errors are minimised. Reprojection errors are calculated at each iteration using the estimated intrinsic and extrinsic camera parameters to reproject observed object points in the image. The reprojected points are used to calculate the reprojection error using the following equation:

$$e_f = \frac{\sum_{i=1}^{n} e_i}{n} = \frac{\sum_{i=1}^{n} \left\|(m_i' - \hat{m}_i')\right\|^2}{n}\,, \tag{15}$$

where e_f is the reprojection error of frame f, e_i is the error of object point i in frame f, m_i' is the object point in the image plane, $\hat{m}_i'$ is the projected point and n is the total number of point correspondences over frame f.

During the stereo calibration process, it was observed that the LMA optimisation algorithm used would become numerically unstable and diverge from a suitable solution. To achieve the best calibration possible, each calibration was attempted multiple times with different termination criteria based upon both the total number of iterations and the change in parameters at which the iterative algorithm stops, ϵ. A calibration image pair was rectified for each calibration and visually inspected to see if horizontal scanlines match correctly. Calibrations that had the lowest reprojection errors and produced good rectified images were selected. For calibration, a planar 8×6 chessboard with square dimensions 80.8×80.8 mm was used during calibration to obtain the intrinsic and extrinsic parameters of the two stereo pairs.

A total of 150 images pairs were taken for both the front and

Source Image | Scale Factor = 2.5 | Scale Factor = 5.0 | Scale Factor = 7.5

Fig. 6. Examples of different scale factors used during perspective undistortion of spherical images. This figure shows that while using a larger scale factor increased the FOV of the images; it also introduces significant distortion around the edge of the image while compression around the centre of the image causes significant loss of detail from the original image.

back stereo pairs. It was found that including all 150 image pairs in the calibration process resulted in unreasonably long computational times and produced calibrations with unacceptable reprojection error. From the 150 image pairs available for each stereo pair, 75 were randomly sampled image pairs were used during the calibrations.

Similar to the approach used in [7], an automatic corner detection algorithm [21] was used. Additionally, sub-pixel accurate corner extraction [31] was used to further refine detected corner locations.

C. Spherical to Planar Projection

The same technique used in [7] is used to implement the longitude-latitude projection described in Section III-E. A projection matrix is defined:

$$[P] = \begin{bmatrix} \frac{|u|}{\theta_u} & 0 & 0 \\ 0 & \frac{|v|}{\phi_v} & 0 \\ 0 & 0 & 1 \end{bmatrix}, \tag{16}$$

where $|u|$ is the image width, $|v|$ is the image height, θ_u is the horizontal FOV of the camera and ϕ_v is the vertical FOV. Since the same cameras that were used in [7] are again used in this project, $\theta_u = \phi_v = \frac{19}{18}\pi$ rad and $|u| = |v| = 640$ pixels, giving the radius of spherical camera model $f_s \approx 193$ px/rad in Eq. (9), (12) and (14).

D. OCamCalib Perspective Undistortion

The coefficients found using the OCamCalib toolbox [30] were used to create lookup tables for applying the perspective undistortion shown in Eq. (11). When performing the undistortion step, a scale factor must be chosen. The scale factor alters the distance between the undistorted image and the camera centre. This causes the scale factor to act as a kind of zoom factor that alters the camera FOV, as shown in Figure 6.

To make the stereo rig as cost-effective as possible, the broadest possible FOV available after the undistortion step is desirable. However, this introduces a significant loss of detail from the centre of the source image. Images are first padded with blank space up to 1280×1280 pixels from 640×640 pixels to prevent loss of detail. The padded images are then undistorted using Eq. (11), which allows for greater detail in the centre of the image due to its larger size, as well as a wider FOV.

E. Stereo Correspondence Algorithms

As covered in Section III-F, to compute the depth of points visible from the stereo rig pixel correspondences in the top and bottom images need to be found. Three separate stereo correspondence algorithms were implemented for this study: Semi-Global Block Matching (SGBM), which is a modified version of the Semi-Global Matching Algorithm (SGM) proposed by Hirschmuller [24]; Semi-Global Block Matching with Weighted Least Squares filtering (SGBM-WLS); and finally PSMNet, which is a pyramid stereo matching network proposed by Chang and Chen [23].

SGBM was implemented for ease of comparison to the previous omnidirectional stereo implementation by Lin and Breckon [7]. The main difference between SGBM and the original SGM algorithm from Hirschmuller [24] is a faster computation time due to only five directions contributing to approximate the global energy function instead of the original eight directions. Aggregated matching costs are calculated by performing line optimisation along each of the five directions and summing for a total matching cost. The final disparity maps are produced by taking the result with the minimum matching cost. This implementation uses the Birchfield-Tomasi sub-pixel metric [32] and some post-processing steps, including a uniqueness check, speckle filtering, and quadratic interpolation.

SGBM-WLS was implemented to improve the quality of disparity maps produced by SGBM, which are prone to noise. The SGBM-WLS filters disparity maps produced using the SGBM algorithm using a Weighted Least Squares (WLS) filter in the form of a Fast Global Smoother [33]. The WLS filter has parameters $\lambda = 800$, and $\sigma = 1.2$ where λ represents the amount of regularization during filtering and sigma is a parameter defining how sensitive the filtering process is to edges in the source image. The parameters used for the SGBM algorithm for both normal SGBM and SGBM-WLS are the same as those used in the OCV-SGBM submission to the KITTI stereo 2015 leaderboards [34] for ease of comparison.

PSMNet [23] is a pyramid stereo matching network consisting of two main modules: spatial pyramid pooling and a 3D CNN. The Spatial Pyramid Pooling (SPP) module allows PSMNet to benefit from global semantic context by using hierarchical context information. The SPP module allows PSMNet to learn the relationship between an object and its sub-regions. A pre-trained PSMNet model, trained on the KITTI 2015

Calibration Method	Average Reprojection Error	Average Reprojection RMS	No. of Iterations	ϵ	No. of Chessboards Detected (front)	No. of Chessboards Detected (back)
Mei's	0.135	0.182	400	1×10^{-7}	36	68
Scaramuzza's SF = 2.5	1.879	1.509	100	0.0001	35	41
Scaramuzza's SF = 5.0	0.557	0.580	100	0.0001	45	51
Scaramuzza's SF = 7.5	0.672	0.713	100	0.0001	33	39

TABLE II. Calibration results using Mei's Calibration Method and Scaramuzza's Calibration Method. Scale factor is denoted by SF. Reprojection errors and RMS values are averaged across the four cameras.

Camera	Reprojection Error	Reprojection RMS	No. of Chessboards Detected
bb	0.189	0.228	41
bf	0.125	0.121	38
tb	0.159	0.143	52
tf	0.137	0.152	48

TABLE III. Calibration results using the ImprovedOCamCalib MATLAB toolbox. This table shows reprojection error and reprojection error rms values for the calibration to find coefficients for Eq. (11).

dataset, was used in this study. When using Scaramuzza's Calibration Method with PSMNet, images were cropped from 1280×1280 to 1280×400 due to limited available GPU RAM.

F. Investigation into the Accuracy of Omnidirectional Stereo

The accuracy of stereo depth sensing using the two proposed calibration methods is evaluated by taking photos of a target at a known distance from the cameras. A total of 67 image pairs were taken at regular intervals in a circle of radius $3m$ around the stereo rig, with 34 images pairs taken for the front cameras and 33 images pairs taken for the back. These image pairs were used to calculate disparity maps using Mei's Calibration Method. Another 67 image pairs were taken with a target at varying points, all with a known distance of $2m$ from the stereo rig in the z-direction to evaluate disparity maps produced using Scaramuzza's Calibration Method at multiple scale factors. Again, 34 images pairs taken for the front cameras and 33 images pairs taken for the back cameras. Disparity maps were generated using SGBM, SGBM-WLS, and PSMNet. The disparity maps were then used to create depth maps, using Eq (13) and (14) for disparity maps produced using Scaramuzza's and Mei's Calibration Method, respectively. The pixel coordinates of the corners of the target were then manually selected from each depth map. The pixel coordinates were used to calculate the mean predicted depth across the target, which was then used to calculate the depth error. As the accuracy of Mei's Calibration Method is expected to vary as points approach the image edges [21], the horizontal distance from the centre of the target to the image centre was also calculated using the target corner pixel coordinates.

V. Results and Discussion

The results of Mei's and Scaramuzza's calibration procedures are presented and discussed in this section. Subsequently, the results from the investigation into the accuracy of omnidirectional stereo are presented and analysed. Finally, achieved FPS throughput using various correspondence algorithms and calibration methods are presented.

A. Calibration

As mentioned in Section III-C, stereo correspondence algorithms rely on rectified images to find correspondences. To successfully rectify images, the estimated intrinsic and extrinsic parameters of the stereo system must be accurate. Any errors in the calibration can cause incorrect matching of pixel correspondences between the images, causing incorrect disparity and therefore incorrect depth estimation. Using the method proposed in Section IV-B, intrinsic and extrinsic parameters were estimated using LMA optimisation, and coefficients for Eq. (11) were found using the ImprovedOCamCalib MATLAB toolbox [30][18].

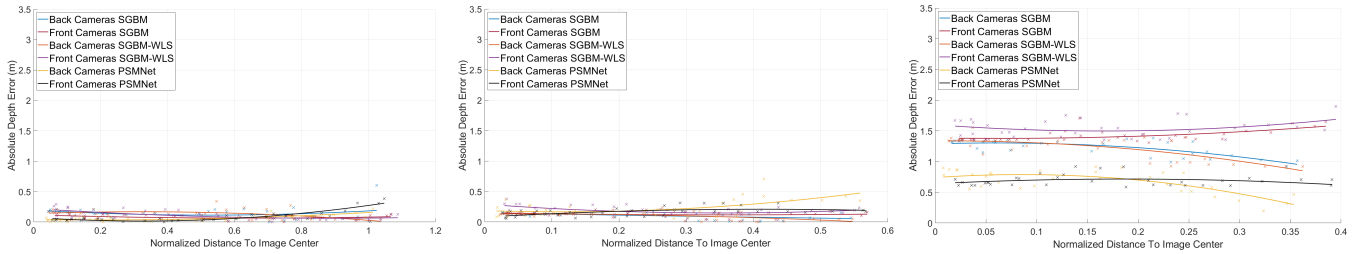
Firstly, the calibration results using the ImprovedOCamCalib MATLAB toolbox are presented in Table III. As mentioned in Section IV-B, multiple stereo calibrations for each of the proposed methods with different termination criteria were performed. The results of the best calibrations are presented in Table II. An RMS value is calculated to evaluate the error in reprojected points, given by:

$$RMS = \sqrt{\frac{\sum_{i=1}^{n} \|(\boldsymbol{m_i'} - \hat{\boldsymbol{m_i'}})\|^2}{2n}}, \tag{17}$$

where $\boldsymbol{m_i'}$ is the point on the image plane, $\hat{\boldsymbol{m_i'}}$ is the projected point on the the image plane and n is the total number of point correspondences over all images.

Table III shows small reprojection errors for all four cameras, with the average result being approximately 0.15 pixels. On average, these results mean that reprojected points after the undistortion step of Eq. (11) are approximately 0.15 pixels away from the actual point. Table III shows similar RMS values to those obtained in [18], which indicates the optimisation process performed by the ImprovedOCamCalib Toolbox has converged successfully. It should be noted that this is not the only step that introduces error into the system, as after Eq. (11) is applied, images still need to be rectified. Any error from the stereo calibration will also impact the quality of rectified images. Table II shows the reprojection errors and RMS values calculated from the performed stereo calibrations.

As seen in Table II, the average reprojection error and RMS when using Mei's Calibration Method are much lower



(a) Scaramuzza's Calibration Method SF = 2.5 **(b)** Scaramuzza's Calibration Method SF = 5.0 **(c)** Scaramuzza's Calibration Method SF = 7.5

Fig. 7. Absolute Depth Error (m) vs Normalized Horizontal Distance to Image Centre using Scaramuzza's Calibration Method at various scale factors. Distance to image centre is normalized by the half the image width, which in this case is 640 pixels.

than that achieved using Scaramuzza's Calibration Method at various scale factors for stereo calibration. Table II shows that in particular, Scaramuzza's Calibration Method used with $SF = 2.5$ achieved poor calibration results.

In reality, overall errors when using Scaramuzza's Calibration Method are increased even further by errors from estimating coefficients for Eq. (11) presented in Table III. Compared to Mei's Calibration Method, the increased re-projection error of Scaramuzza's Calibration Method means that disparity values estimated using Scaramuzza's Calibration Method are more likely to be incorrect than when using Mei's Calibration Method.

The extrinsic parameters for the front stereo pair using Mei's and Scaramuzza's Calibration Methods ($SF = 5.0$) were estimated to be,

$$R_M = \begin{bmatrix} 0.9999 & -0.0107 & 0.0093 \\ 0.0107 & 0.9999 & -0.0003 \\ -0.0093 & 0.0004 & 0.9999 \end{bmatrix} \quad T_M = \begin{bmatrix} 7.077 \\ -181.1 \\ -1.080 \end{bmatrix} mm$$

$$R_S = \begin{bmatrix} 0.9999 & -0.0144 & 0.0043 \\ 0.0145 & 0.9999 & -0.0068 \\ -0.0042 & 0.0069 & 0.9999 \end{bmatrix} \quad T_S = \begin{bmatrix} 5.058 \\ -183.8 \\ -2.782 \end{bmatrix} mm$$

respectively. The two sets of estimated extrinsic parameters, although similar, do have differences due to errors in the calibration process. The two translation vectors T_M and T_S give baselines $b_M = 181.2mm$ and $b_S = 183.8mm$ respectively. The rotation matrices R_M and R_S show that the stereo rig is set up so that the two cameras optical axis are already well aligned, as $R_M \approx R_S \approx I$, where I is the identity matrix.

B. Experimental Results

The results of the experiment described in Section IV-F are presented here. As mentioned in Section III-E, the absolute depth error is expected to increase as image points get closer to the epipoles when using a spherical camera model. After the longitude-latitude projection is performed on a spherical image, the epipoles are placed at the vertical edges of the image. This prediction is confirmed by the results shown in Figure 8, where absolute depth error is shown to increase as image points get further from the centre of the image and head towards the epipoles. The error around the image centre is acceptable; however, the absolute depth error increases significantly towards the edge of the images. This trend appears for all of the stereo correspondence algorithms used, for both hand-crafted algorithms such as SGBM [24] and deep-learning-based approaches like PSMNet [23].

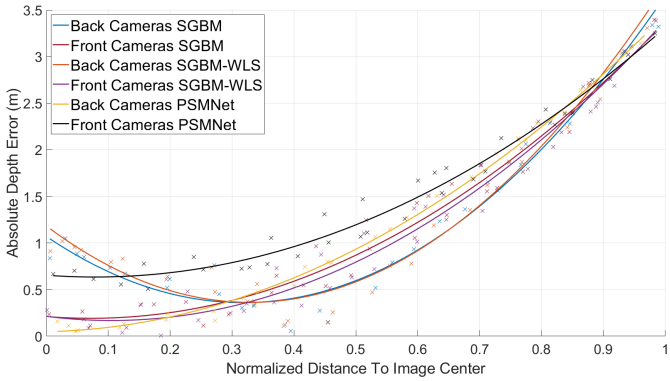


Fig. 8. Absolute Depth Error (m) vs Normalized Horizontal Distance to Image Centre using Mei's Calibration Method. Distance to image centre is normalized by the half the image width, which in this case is 320 pixels.

The line of best fit obtained from measurements of depth error while using Mei's Calibration Method with SGBM-WLS, shown in Figure 8, is overlayed onto a disparity map to show how depth accuracy varies across a disparity map when using a spherical camera model, and is shown in Figure 9.

The overall calibration error is acceptable when using Mei's Calibration Method, and as such, the depth error shown in Figure 8 is almost entirely caused by the points appearing close to the epipoles. Such a high level of inaccuracy towards the edges of the images, as seen in Figure 9, means that although the use of a fish-eye lens and spherical camera model allows for disparity information to be calculated for a large

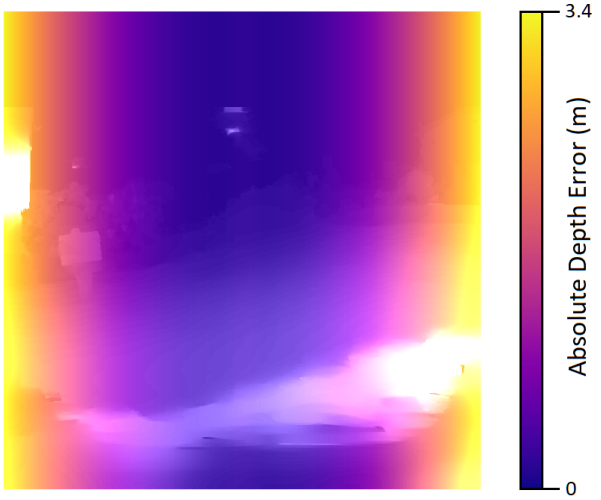


Fig. 9. Absolute Depth Error (m) overlayed onto a disparity map generated using Mei's Calibration Method with SGBM-WLS.

Rotated Source Image

Mei's Calibration Method
Rectified Image

Scaramuzza's Calibration Method
SF = 5.0 Rectified Image

SGBM - Scaramuzza's Calibration Method
SF = 5.0

SGBM - WLS Mei's
Calibration Method

PSMNet - Scaramuzza's Calibration Method
SF = 5.0

Fig. 10. Disparity map pipeline. This figure shows one image from an image pair taken from the experiment outlined in Section IV-F which is then rectified using the two proposed calibration techniques. The image pairs are then used to create disparity maps which are shown here. The PSMNet - Scaramuzza's Calibration Method disparity map has been cropped to show the region of interest (the target). Scale factor is denoted by SF. Disparity maps shown above have been converted from greyscale using a plasma colormap for clarity.

FOV, a significant portion of the generated disparity map is useless. This renders the perceived advantage that using a fish-eye lens with a spherical camera model will allow for a disparity maps with greater than $180°$ FOV untrue as there are effective blindspots. Results shown in Figure 7 show disparity maps produced using Scaramuzza's Calibration Method do not encounter increasing depth error as image points appear further from the image centre, as experienced when using a spherical camera model. Comparing Figure 7(c) to Figures 7(a) and 7(b) shows when using a large scale factor significant depth error is introduced, caused by increasing distortion of rectified images, as illustrated in Figure 6. Although, as presented in Table II, the reprojection error of calibrations produced using Scaramuzza's Calibration Method are greater than that of Mei's Calibration Method, by comparing Figures 7 and 8 it is clear that using Scaramuzza's Calibration Method with a suitable scale factor achieves significantly improved depth accuracy across virtually all of the produced disparity map, in particular at the image edges.

C. Real-time Omnidirectional Stereo

The FPS of different combinations of calibration methods and stereo correspondence algorithm are presented in Table IV. By performing image undistortion, spherical projection and the SGBM algorithm on a GPU, this paper achieves a higher FPS of up to 14.4 FPS than achieved in previous work by Lin and Breckon [7]. Mei's Calibration Method achieved greater FPS due to it using lower resolution rectified images 640×640 than Scaramuzza's Calibration Method, which used 1280×1280 images.

Matching Algorithm	FPS
SGBM	0.47
SGBM-WLS	0.22
SGBM-CUDA	5.82
PSMNet	0.25

Scaramuzza's Calibration Method

Matching Algorithm	FPS
SGBM	2.05
SGBM-WLS	1.01
SGBM-CUDA	14.4
PSMNet	0.34

Mei's Calibration Method

TABLE IV. FPS of disparity maps achieved when using Scaramuzza's and Mei's Calibration Method. When using different scale factors for Scaramuzza's Calibration Method the FPS remained the same.

Correspondence Algorithm	Calibration Method	
	Mei's	Scaramuzza's
SGBM	74.651	85.979
SGBM-WLS	73.202	86.986
PSMNet	99.997	100.00

TABLE V. **Density of coverage in disparity maps produced using Mei's and Scaramuzza's Calibration Methods.** Values are averaged over the different scale factors used when using Scaramuzza's Calibration Method as similar densities were achieved.

The achieved density of coverage in disparity maps generated for the experiment proposed in Section IV-F is presented in Table V. This table shows the average percentage of

matched pixels in disparity maps using Mei's and Scaramuzza's Calibration Methods with SGBM, SGBM-WLS and PSMNet. Table V shows SGBM and SGBM-WLS achieve similar performance in terms of coverage, while PSMNet outperforms both methods, achieving 100% coverage.

Some of the disparity maps generated for the experiment proposed in Section IV-F are shown in Figure 10. In these disparity maps, details such as the experiment target are clearly visible. However, disparity maps often contain noise caused by the automatic shifting of ISO sensitivity and white balance settings done by the two Ricoh Theta S cameras. This problem was also experienced by Lin and Breckon [7], who used the same omnidirectional cameras used in this study.

VI. Conclusion

In conclusion, this paper successfully demonstrates an omnidirectional stereo system that produces complete 360° FOV disparity maps using consumer-grade hardware. A new camera configuration is adopted to overcome issues highlighted in [9], and better performance is achieved than [7] by using GPU computing to run correspondence algorithms, producing complete 360° disparity maps at up to 14.4 FPS. By using more recent stereo correspondence algorithms, greater disparity coverage is achieved than [7]. Furthermore, the accuracy of an omnidirectional stereo system such as [7] using the spherical camera model proposed by Mei and Rives [8] is investigated, revealing that it suffers from poor depth accuracy as the angle from the cameras optical axis approaches the limit of the cameras field of vision. This paper overcomes this problem by using the perspective undistortion function proposed by Scaramuzza et al. [10] to produce a stereo system with improved depth accuracy at large angles from the optical axis than [7]. Further work could improve upon this system by making use of higher resolution cameras, as well as improved computing hardware. This system would benefit from testing in real-life scenarios on the road, where LiDAR sensors could be used to obtain ground-truth measurements to evaluate the system's accuracy across an entire scene. This system is not only valuable for driverless vehicles but has broader applications such as use in unmanned aerial vehicles, object detection and 3D scene reconstruction.

References

[1] National Highway Traffic Safety Administration, "Automated Vehicles for Safety," *Nat. Highway Traffic Saf. Admin.*, 2018.

[2] J. Rios-Torres and A. A. Malikopoulos, "Impact of Partial Penetrations of Connected and Automated Vehicles on Fuel Consumption and Traffic Flow," *IEEE Trans. Intell. Vehicles*, vol. 3, no. 4, pp. 453–462, 2018.

[3] J. B. Greenblatt and S. Saxena, "Autonomous taxis could greatly reduce greenhouse-gas emissions of US light-duty vehicles," *Nature Climate Change*, vol. 5, no. 9, pp. 860–863, 2015.

[4] Daimler, "Mercedes-Benz Intell. Drive: The Intell. car," Accessed on: Dec. 21, 2020. [Online]. Available: https://media.daimler.com/marsMediaSite/en/instance/ko/Mercedes-Benz-Intell.-Drive-The-Intell.-car.xhtml?oid=9904196

[5] I. Spectrum, "How Google's Self-Driving Car Works," Accessed on: Dec. 21, 2020. [Online]. Available: https://spectrum.ieee.org/automaton/robotics/artificial-Intell./how-google-self-driving-car-works

[6] Velodyne, "Velodyne HDL-64E LiDAR specification sheet," 2020, Accessed on: Dec. 21, 2020. [Online]. Available: https://velodynelidar.com/products/hdl-64e/

[7] K. Lin and T. Breckon, "Real-time low-cost omni-directional stereo vision via bi-polar spherical cameras," in *Proc. Int. Conf. Image Anal. Recognit.* Springer, June 2018, pp. 315–325.

[8] C. Mei and P. Rives, "Single view point omnidirectional camera calibration from planar grids," in *Proc. - IEEE Int. Conf. Robot. and Automat.*, 2007, pp. 3945–3950.

[9] K. Lin, "Omnidirectional Stereo Vision for Environment Mapping in Future Driver-less Vehicles," Master's thesis, Durham Univ. Dept. of Eng., 2017.

[10] D. Scaramuzza, A. Martinelli, and R. Siegwart, "A flexible technique for accurate omnidirectional camera calibration and structure from motion," in *Proc. 4th IEEE Int. Conf. Comput. Vision Syst.*, 2006, pp. 45–45.

[11] Bosch, "Fourth generation long-range radar sensor (LRR4) Specification Sheet," Accessed on: Dec. 21, 2020. [Online]. Available: https://cds.bosch.us/themes/bosch_cross/amc_pdfs/LRR4_292000P0ZH_EN_low.pdf

[12] StereoLab, "StereoLab Zed 2," Accessed on: Dec. 21, 2020. [Online]. Available: https://store.stereolabs.com/products/zed-2?_ga=2.253170889.137628384.1608580171-1038765621.1608580171

[13] Toposens, "Toposens TS3 Datasheet," 2020, Accessed on: Dec. 21, 2020. [Online]. Available: https://toposens.com/wp-content/uploads/2020/08/TS3_Datasheet_V1.1.pdf

[14] M. Merano, "Tesla's new patent shows path to Elon Musk's pure vision FSD approach," 2021, Accessed on: Feb. 4, 2021. [Online]. Available: https://www.teslarati.com/tesla-pure-vision-fsd-patent-elon-musk/

[15] L. Puig, J. Bermúdez, P. Sturm, and J. J. Guerrero, "Calibration of omnidirectional cameras in practice: A comparison of methods," *Comput. Vis. Image Understanding*, vol. 116, pp. 120–137, 2012.

[16] Ricoh, "Ricoh Theta S - User Guide," 2015, Accessed on: March. 22, 2021. [Online]. Available: https://support.theta360.com/uk/manual/s/

[17] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 11, pp. 1330–1334, 2000.

[18] S. Urban, J. Leitloff, and S. Hinz, "Improved wide-angle, fisheye and omnidirectional camera calibration," *ISPRS J. Photogrammetry Remote Sens.*, vol. 108, pp. 72–79, 2015.

[19] C. Geyer and K. Daniilidis, "A unifying theory for central panoramic systems and practical implications," in *Eur. Conf. Comput. Vis.* Springer, 2000, pp. 445–461.

[20] J. P. Barreto and H. Araujo, "Issues on the geometry of central catadioptric image formation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2, 2001, pp. II–II.

[21] B. Li, L. Heng, K. Koser, and M. Pollefeys, "A multiple-camera system calibration toolbox using a feature descriptor-based calibration pattern," in *IEEE Int. Conf. Intell. Robots Syst.*, 2013, pp. 1301–1307.

[22] M. Menze and A. Geiger, "Object scene flow for autonomous vehicles," in *Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 3061–3070.

[23] J.-R. Chang and Y.-S. Chen, "Pyramid stereo matching network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 5410–5418.

[24] H. Hirschmüller, "Stereo processing by semiglobal matching and mutual information," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 328–341, 2008.

[25] W. Gao and S. Shen, "Dual-fisheye omnidirectional stereo," in *IEEE Int. Conf. Intell. Robots Syst.*, 2017, pp. 6715–6722.

[26] C. Won, J. Ryu, and J. Lim, "OmniMVS: End-to-end learning for omnidirectional stereo matching," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 8987–8996.

[27] C. Ma, L. Shi, H. Huang, and M. Yan, "3D Reconstruction from Full-view Fisheye Camera," *arXiv preprint arXiv:1506.06273*.

[28] J. Heikkila and O. Silven, "Four-step camera calibration procedure with implicit image correction," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 1997, pp. 1106–1112.

[29] *The OpenCV Reference Manual*, 4th ed., OpenCV, March 2021.

[30] D. Scaramuzza, A. Martinelli, and R. Siegwart, "A toolbox for easily calibrating omnidirectional cameras," in *IEEE Int. Conf. Intell. Robots Syst.*, 2006, pp. 5695–5701.

[31] W. Förstner and E. Gülch, "A Fast Operator for Detection and Precise Location of Distinct Points, Corners and Centres of Circular Features," in *ISPRS Intercommission Workshop*, 1987, pp. 281–305.

[32] S. Birchfield and C. Tomasi, "A pixel dissimilarity measure that is insensitive to image sampling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 4, pp. 401–406, 1998.

[33] D. Min, S. Choi, J. Lu, B. Ham, K. Sohn, and M. N. Do, "Fast global image smoothing based on weighted least squares," *IEEE Trans. Image Process.*, vol. 23, no. 12, pp. 5638–5653, 2014.

[34] OpenCV, M. Menze, and A. Geiger, "OpenCV Semi-Global Block Matching [OCV-SGBM] - KITTI," Accessed on: Dec. 26, 2020. [Online]. Available: http://www.cvlibs.net/datasets/kitti/eval_scene_flow_detail.php?benchmark=stereo&result=67877bd6fcc4316342 1fa0108c7df83bbc69fea3