Extrinsics Self-calibration of the Surround-view System: A Weakly Supervised Approach

Yang Chen, Ying Shen, Lin Zhang, *Senior Member, IEEE,* Tianjun Zhang, and Yicong Zhou, *Senior Member, IEEE*

***Abstract*—An SVS usually consists of four wide-angle fisheye cameras mounted around the vehicle to sense the surrounding environment. From the images synchronously captured by all cameras, a top-down surround-view can be synthesized, on the premise that both intrinsics and extrinsics of the cameras have been calibrated. At present, the intrinsics calibration approach is relatiavely complete and can be pipelined, while the extrinsics calibration is still immature. On one hand, the existing manual calibration schemes are usually reliable, but need to be conducted by professionals in specific sites, which is undoubtedly cumbersome. On the other hand, most of the existing self-calibration schemes are based on the low-level features and their stability and robustness are usually unsatisfactory. As far as we know, an effective extrinsics self-calibration scheme designed specially for the SVS is still lacking. To fill such a research gap to some extent, we propose a novel self-calibration scheme which follows a weakly supervised framework, namely WESNet (Weakly-supervised Extrinsics Self-calibration Network). The training of WESNet consists of two stages. First, we utilize the corners in a few calibrated images as the weak supervision to roughly optimize the network by minimizing the re-projection loss. Then, after the convergence in the first stage, we additionally introduce a self-supervised photometric loss term that can be constructed only by the information from the natural images for further fine-tuning. Besides, to support training, we totally collected 19,078 groups of synchronously captured fisheye images under various environmental conditions. To our knowledge, thus far this is the largest surround-view dataset containing original fisheye images. By means of learning prior knowledge from the training data, WESNet takes the original fisheye images synchronously collected as the input, and directly yields extrinsics end-to-end with little labor cost. Its efficiency and efficacy have been corroborated by extensive experiments conducted on our collected dataset. To make our results reproducible, source code and the collected dataset have been released at https://github.com/xiaofeng94/RefineDNet-for-dehazing.**

***Index Terms*—Surround-view system, weakly supervised learning, extrinsics calibration, photometric loss, surround-view dataset.**

I. Introduction

AS an indispensable component of modern ADAS [1], the surround-view system (SVS) has been installed by more and more vehicles. In the SVS, four wide-angle fisheye cameras that cover the 360-degree field of view around the vehicle are mounted. Based on the multi-view geometry knowledge [2], the top-down surround-view images can be synthesized at run time from the multi-stream video collected synchronously. With the surround-view, the driver can intuitively check whether there are obstacles around the vehicle without blind spots and grasp their relative orientations and distances. In this way, the occurrence of scraping, collision and other accidents can be avoided effectively. Besides, the surround-view also plays an important role in multiple computer vision tasks [3] towards driving assistance, such as the parking-slot detection [4], [5], autonomous parking [6], [7], pedestrian detection [8], [9] and so on.

To synthesize high-quality surround-views, the accurate intrinsics and extrinsics calibration of cameras in the SVS are necessary. At present, the performance of intrinsics calibration schemes is relatively satisfactory. The manufactures can complete the cameras' manufacturing and intrinsics calibration in a streamlined manner. Besides, since the cameras are always tightly encapsulated, the intrinsics will usually remain fixed after being produced. Therefore, the intrinsics re-calibration is not frequently required in most cases. Relatively speaking, the techniques for extrinsics calibration are still not mature yet. The existing schemes for extrinsics calibration mainly fall into two categories, manual ones [10]–[15] and self-calibration ones [16]–[21], and their limitations are mainly manifested in the following two aspects:

1) Although the existing manual calibration schemes often preform reliably, most of them are cubersome and labor-consuming. When these methods are utilized, the vehicle needs to be driven by professionals to a specific calibration site, and then the calibration can be completed using the patterns with the known-scale regularly arranged in the site. It can be seen that except for the high labor cost, these methods also have specific restrictions on the working environment. As aforementioned, due to collisions or bumps, sometimes the extrinsics of the SVS may change, which leads to the result that the manual schemes can only work with the professional assistance in an offline manner, but are not applicable for the online environment during driving.
2) The existing self-calibration schemes generally have obvious limitations with respect to the robustness and

This work was supported in part by the National Natural Science Foundation of China under Grants 61973235, 61936014 and 61972285, in part by the Natural Science Foundation of Shanghai under Grant 19ZR1461300, and in part by the Shanghai Science and Technology Innovation Plan under Grant 20510760400. (*Corresponding author: Lin Zhang.*)

Yang Chen, Ying Shen, Lin Zhang and Tianjun Zhang are with the School of Software Engineering, Tongji University, Shanghai 201804, China (email: {2011439, yingshen, cslinzhang, 1911036}@tongji.edu.cn).

Yicong Zhou is with the Department of Computer and Information Science, University of Macau, Macau 999078, China (e-mail: yicongzhou@um.edu.mo).

the stability. This is because most of them only adopt the low-level geometric features on the ground, such as pixels, points, and lines, and estimate the extrinsics via the relatively ideal mathematical model. On the one hand, such low-level features are sensitive to the natural noise and the accuracy of the system will evidently decline in nonideal environments. On the other hand, some features (such as lines) are not widely available in the natural environment. Without required features on the ground, the corresponding self-calibration method will fail.

On account of the limitations aforementioned, as far as we know, there is still no existing extrinsics self-calibration scheme specially designed for the SVS that can be stably applicable in various environments. In most commercial solutions, drivers have to drive to 4S stores for calibration or re-calibration by professionals. This is undoubtedly troublesome for both customers and automobile manufacturers. Thus, many manufacturers are now looking for effective self-calibration schemes. To fill in this research gap to some extent, we propose a novel weakly-supervised scheme towards the extrinsics self-calibration of the SVS. In summary, our contributions are mainly threefolds:

1) A weakly supervised network [22] for extrinsics calibration of the SVS, namely WESNet, is proposed. Based on the prior knowledge learned from the training data, WESNet can yield the extrinsics of the SVS in an end-to-end manner with the input of original fisheye images. Since it is difficult to obtain the accurate extrinsics as the ground truth (GT), we do not directly label with the GT extrinsics and conduct fully supervised learning, but follow a weakly supervised framework. Specifically, the corner information in the calibrated images (collected over the calibration site) is taken as the geometric supervision, and in the first stage, the network is optimized by minimizing the re-projection loss.
2) A novel photometric loss as self-supervised information [23] is designed, so as to mine more supervision information from the training images themselves. Inspired by Zhang *et al.*'s scheme in [21], OECS, we model the imaging discrepancy in the common-view regions of adjacent cameras in the SVS as the photometric loss, and expect to minimize the loss as much as possible in the training process, so as to synthesize seamless and high-quality surround-views. When the training fully based on the geometric supervision converges, the self-supervised photometric loss will be introduced to fine-tune the network to improve the estimation accuracy.
3) To facilitate the study of the extrinsics calibration or any other computer vision tasks based on the surround-views, we collected a large-scale surround-view dataset covering a variety of environmental conditions. Such a dataset contains 19,078 groups of high-resolution fisheye-images and the corresponding synthesized surround-views under different environmental conditions, covering the ground with several kinds of lane-lines and tiles, the cement road, the narrow path, and the road exposed to strong sunlight. Besides, a data augmentation method based on the homography transformation [24] is also proposed, for the sake of improving the richness of the extrinsics of collected data. It is worth mentioning that, to our knowledge, this is the largest surround-view dataset containing original fisheye images. To make our results reproducible, source code and the collected dataset in this paper are online available at ...

The remainder of this paper is organized as follows. Section II introduces related studies. Section III makes an overview of the imaging principle of the SVS. Section IV and V present our proposed network, WESNet, and the collected dataset in detail, respectively. Experimental results are reported in Section VI. Finally, Section VII concludes the paper.

II. Related Work

A. Extrinsics Calibration of the Surround-View System

To synthesize the surround-view image, the SVS needs to be both intrinsically and extrinsically calibrated. Since the intrinsics calibration is relatively mature and can satisfy the industrial requirements in most cases, in this paper, we mainly focus on the aspect of extrinsics. Based on whether the calibration can be conducted automatically, existing extrinsics calibration methods are mainly divided into two categories, the manual ones and the self-calibration ones.

Manual calibration methods. In manual calibration methods, specific patterns, such as corners, circles or lines, are necessary so as to offer the reference information. These patterns are usually repeatedly and equidistantly printed on the calibration site or some portable reference targets like the chessboard, and the coordinate of each pattern in the world coordinate system can be easily obtained. The driver needs to park the vehicle equipped with the SVS at the appropriate position, and then captures the calibrated images. After that, the extrinsics can be solved by establishing the mapping relationships of patterns between the pixel coordinates and the world coordinates.

In [10], Liu *et al.* firstly proposed the basic theoretical model of the SVS, pointing out that the mapping relationship between the undistorted fisheye image and the surround-view can be determined by homography estimation. However, the author did not offer a specific calibration pipeline. In fact, as far as we know, at that time, the SVS was still in the early stage where the tiles of fixed size on the ground are usually considered as the simple calibration patterns, which is undoubtedly unsatisfactory in the accuracy. In [11], Hedi *et al.* presented a two-stage offline calibration pipeline. In its first stage, the vehicle was parked on the calibration site filled with the chessboard markers. Then the extrinsics were roughly estimated via the homography estimation. After that, an optimization approach to minimize the stitching loss was conducted, which was also the second stage of the pipeline. Zhang *et al.*'s solution proposed in [12] is a calibration-chart-based approach, which utilized Harris corners [25] and BRIEF descriptors [26] to find paired features between the calibration-chart and the collected calibrated images. One eminent feature of their work is that except for the geometric

alignment, photometric alignment is also introduced. In the scheme presented by Shao *et al.* [13], instead of driving the vehicle to a fixed position in a specific calibration site, a single chessboard is the only demand. And a novel refinement procedure that jointly optimized camera poses in a closed-loop manner was adopted. In recent years, a new variant of the SVS, 3D SVS, has attracted a lot of research interest and the studies on extrinsics calibration of the 3D SVS naturally emerged, such as Gao *et al.*'s work [14] and Zhang *et al.*'s one [15]. Actually, these methods are not significantly different from the aforementioned schemes designed for the conventional SVS in the calibration aspect. Specifically, Gao *et al.*'s solution is based on the calibration site printed with chessboard markers, and Zhang *et al.*'s one relies on the calibration chessboard, which are similar to the design in [11] and [13], respectively.

With the assistance of calibration patterns, manual calibration methods often perform satisfactorily in both the stability and the accuracy. However, such calibration approaches usually need to be operated by professionals in specific sites, which causes high cost of manpower and materials. Besides, since these methods can only be applicable to the off-line environment, once the camera poses in the SVS changes due to collisions or bumps, the extrinsics obtained by the initial manual calibration will become inaccurate, and obvious geometric misalignment will appear in the synthesized surround-view.

Self-calibration methods. The self-calibration schemes are independent of the specific calibration pattern, and can recover the extrinsics only from the images taken in natural scenes, which effectively reduce the labor cost. In [16], Zhao *et al.* first detected multiple vanishing points of lane markings on the road via the weighted least squares method, and then with the estimated vanishing points, the pose of the multi-camera system relative to the world coordinate system was solved. In [17], Choi *et al.* also designed a lane-line based extrinsics self-calibration pipeline for the surround-view case, in which the SVS was calibrated by aligning lane markings across images of adjacent cameras. It can be seen that the aforementioned self-calibration frameworks both made an assumption for the target application environment, that is, there must be two parallel lane-lines clearly observed in the field of view. However, this is an assumption that cannot usually be satisfied. For example, lane-lines on the ground may be crooked and faded, or the car is likely to run on a rural path without lane-lines. Heng *et al.* [18], [19] resorted to visual SLAM systems to calibrate the extrinsics of the SVS and proposed an infrastructure-based pipeline. In their pipeline, there are no specific limitations on the application scope; however, the vehicle equipped with the SVS needs to travel in the calibration area for a while to establish the map, which is quite a time-consuming approach and unlikely to satisfy the industrial portability requirement. As far as we know, the only two existing relatively lightweight self-calibration schemes which are applicable to the SVS are Liu *et al*.'s method [20] and Zhang *et al.*'s [21]. They all deeply dissected the online extrinsics correction problem and offered effective solutions. In [20], Liu *et al*. proposed two models, namely the "Ground Model" and the "Ground-Camera Model", and both of them can correct extrinsics by minimizing photometric errors with the steepest descent [27]. In [21], Zhang *et al.* designed a novel model, bi-camera model, to construct the least-square errors [28] on the imaging planes of two adjacent cameras and then optimize camera poses by the LM (Levenberg-Marquardt) scheme [29]. Since they focused on the "online correction" rather than the "calibration", a rough initial extrinsics needs to be offered as the input.

At present, most of the existing self-calibration methods can only utilize the low-level features, such as pixels, keypoints [30]–[32] or lines, to solve the extrinsics by aligning the features on different views. As discussed in Sect. I, these methods usually perform satisfactorily in ideal environments, but may fail without required textures on the ground. Compared with them, the solution proposed in this work follows a weakly supervised learning framework. Without any prior, our designed network can effectively extract deep-level features and yield the accurate extrinsics end-to-end.

B. Learning-based Calibration of the Camera System

In recent years, deep learning has shown the superior performance in various computer vision tasks. Towards the calibration problem of camera systems, which is a classical visual task, more and more learning-based solutions were proposed. In [33], Workman *et al.* proposed to regress the intrinsics of the camera directly from a single-shot via a convolutional neural network (CNN), namely FocalNet. Giering *et al.*'s approach in [34] is also an end-to-end CNN-based scheme, which takes a multi-modal input including the point clouds from lidar, the optical flow maps and the RGB images. By solving a 9-class classification problem where each class corresponds to a particular x-y shift on an ellipse, the real-time lidar-video registration can be realized. In [35], Schneider *et al*. designed a new network named RegNet firstly towards the extrinsics calibration of the lidar-camera system. Since it doesn't take geometric relationships into account, it has to be retrained each time the sensor intrinsics change. In contrast, the method presented in [36], namely CalibNet, solves the problem in a weakly-supervised manner by attempting to reduce the dense photometric error and point cloud distance error between the misaligned and the target depth maps. Despite some learning-based calibration schemes have been proposed, as far as we know, they are all not applicable to the surround-view case. Besides, most of the existing schemes are fully supervised, and the GTs are from traditional offline calibration solutions, implying the accuracy of the trained networks is limited. For the consideration of the aforementioned limitations, our proposed WESNet, which is specially designed for the SVS, follows a weakly supervised framework. During its training approach, rather than generating GTs via existing offline calibration schemes, we take the reprojection loss of corners on the calibrated images as the weak supervision information. In addition, a novel photometric loss is also introduced as the self-supervision information to further improve the performance of the network.

III. Overview of the Surround-view System

This section describes the imaging process of surround-view system, specifically, how to generate a surround-view from

images captured by the cameras mounted around the vehicle. To synthesize a surround-view image, the mapping relationship of a point between the pixel coordinate on the original fisheye image and that on the surround-view should be established. Since the relationship is relatively complex in form, we divide it into two parts, the mapping relationship from the pixel coordinate of the fisheye image to the ground coordinate and that from the 3D ground coordinate to the pixel coordinate in the bird's-eye-view. Next, we will introduce these two parts in detail.

Given the ground coordinate system O_G and a four-camera SVS (cameras are represented as C_1, C_2, C_3, C_4), the poses of cameras in O_G are denoted by $\boldsymbol{T}_{C_1G}$, $\boldsymbol{T}_{C_2G}$, $\boldsymbol{T}_{C_3G}$, $\boldsymbol{T}_{C_4G}$, respectively. The pose matrix $\boldsymbol{T}_{C_iG}$ is 4×4 and of 6 DOF (Degrees of Freedom), which can be expressed as,

$$\boldsymbol{T}_{C_iG} = \begin{bmatrix} \boldsymbol{R}_i & \boldsymbol{t}_i \\ \boldsymbol{0}^T & 1 \end{bmatrix}, i = 1, 2, 3, 4 \tag{1}$$

where $\boldsymbol{R}_i$ is an orthonormal 3×3 rotation matrix while $\boldsymbol{t}_i$ is a three dimensional translation vector.

For the transformation from the ground coordinate system to the pixel coordinate system, we formulate it with the pinhole camera model. Given an arbitrary point in the ground coordinate system $\boldsymbol{P}_G = [X_G, Y_G, Z_G, 1]^T$ in O_G, its corresponding pixel coordinate $\boldsymbol{p}_{C_i}$ in the camera coordinate system of C_i is given by,

$$\boldsymbol{p}_{C_i} = \frac{1}{Z_{C_i}} \boldsymbol{K}_{C_i} \boldsymbol{T}_{C_iG} \boldsymbol{P}_G, i = 1, 2, 3, 4 \tag{2}$$

where Z_{C_i} is the depth of $\boldsymbol{P}_G$ in camera C_i's coordinate system, and $\boldsymbol{K}_{C_i}$ is the 3×3 intrinsic matrix of camera C_i, which can be estimated together with the distortion coefficient matrix by Zhang's salient work [37] and some subsequent work of others [38], [39]. Concretely, the form of $\boldsymbol{K}_{C_i}$ is

$$\boldsymbol{K}_{C_i} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \tag{3}$$

where f_x, f_y, c_x, c_z are all camera intrinsic parameters. And it's worth mentioning that $\boldsymbol{p}_{C_i}$ is an undistorted point.

Compared with the above model, the transformation from the bird's-eye-view coordinate system to the ground coordinate system is much simpler, which is essentially a similarity transformation. The bird's-eye-view image can be generated by projecting a camera image to the ground, namely the plane $Z_G = 0$ in O_G. For a point $\boldsymbol{p}_G = [u_G, v_G, 1]^T$ in the bird's-eye-view coordinate system whose corresponding point in the ground coordinate system is $\boldsymbol{P}_G$, the transformation between them is given as,

$$\begin{bmatrix} u_G \\ v_G \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{d_{X_G}} & 0 & \frac{W}{2d_{X_G}} \\ 0 & -\frac{1}{d_{Y_G}} & \frac{H}{2d_{Y_G}} \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X_G \\ Y_G \\ 1 \end{bmatrix} \tag{4}$$

where X_G, Y_G, Z_G are the coordinate values of $\boldsymbol{P}_G$, and d_{X_G} and d_{Y_G} are the size of each pixel[1], W and H are the width and height of the scope covered by the surround-view image. It is worth mentioning that because $Z_G = 0$, it is ignored implicitly here. Denote the transformation matrix from $\boldsymbol{P}_G$ to $\boldsymbol{p}_G$ by $\boldsymbol{K}_G$, and then Eq. 4 can be simplified as,

$$\boldsymbol{p}_G = \boldsymbol{K}_G \boldsymbol{P}_G \tag{5}$$

By combining Eq. 2 and Eq. 5, we can get,

$$\boldsymbol{p}_{C_i} = \frac{1}{Z_{C_i}} \boldsymbol{K}_{C_i} \boldsymbol{T}_{C_iG} \boldsymbol{K}_G^{-1} \boldsymbol{p}_G \tag{6}$$

With Eq. 6, we are able to establish a complete mapping between $\boldsymbol{p}_G$ from ground coordinate system and $\boldsymbol{p}_{C_i}$ from bird's-eye-view coordinate system. With the injective relationship, considering each point $\boldsymbol{p}_G$ in the bird's-eye-view image $\boldsymbol{I}_{GC_i}$ captured by camera C_i, the corresponding pixel value is represented as,

$$\boldsymbol{I}_{GC_i}(\boldsymbol{p}_G) = \boldsymbol{I}_{C_i}(\boldsymbol{p}_{C_i}) \tag{7}$$

where $\boldsymbol{I}_{C_i}$ is the undistorted image captured by camera C_i. Mapping the images captured by cameras C_1, C_2, C_3, C_4 to bird's-eye views and then stitching them appropriately, a complete surround-view image can be synthesized.

IV. Weakly Supervised Camera Extrinsic Estimation

With accurate extrinsics, seamless surround-view images can be synthesized at run-time. However, as discussed in Sect. I, existing manual calibration schemes are usually laborious so that they can't be applied in the online manner, and self-calibration schemes perform unsatisfactorily in the robustness and the generalization. To provide a robust and lightweight solution for the extrinsics calibration of SVS, in this paper, we proposed a learning-based solution following the weakly supervised framework for camera extrinsics' estimation. Such a scheme is based on an end-to-end lightweight CNN, namely WESNet, which can yield the extrinsics directly from four input fisheye images captured synchronously by the cameras mounted around the vehicle. Under the weakly supervised framework, we mainly leverage the re-porjection loss of the corners of the calibration site instead of labelling every image in the dataset with its corresponding GT extrinsics.

Training with the weak supervision information, WESNet can offer a rough extrinsics estimation but the accuracy is still insufficient. To mine more image information from the training datas themselves, we also introduce the self-supervised photometric loss to fine-tune the network after the weakly supervised loss converges. Thus, the accuracy of the network can be further improved so as to synthesize seamless surround-views.

A. Network Architecture

The basic architecture of WESNet is designed mainly following the advice of regression frameworks for calibration [35], [36] as well as the common knowledge in CNN area. Fig. 1 illustrates the configurations of our network. Here, a residual bottleneck [40] with a 3×3 convolution followed by a 1×1 one is used as the basic building block.

[1]More accurately, each pixel in the surround-view image corresponds to a $d_{X_G} \times d_{Y_G}$ physical area on the ground plane.

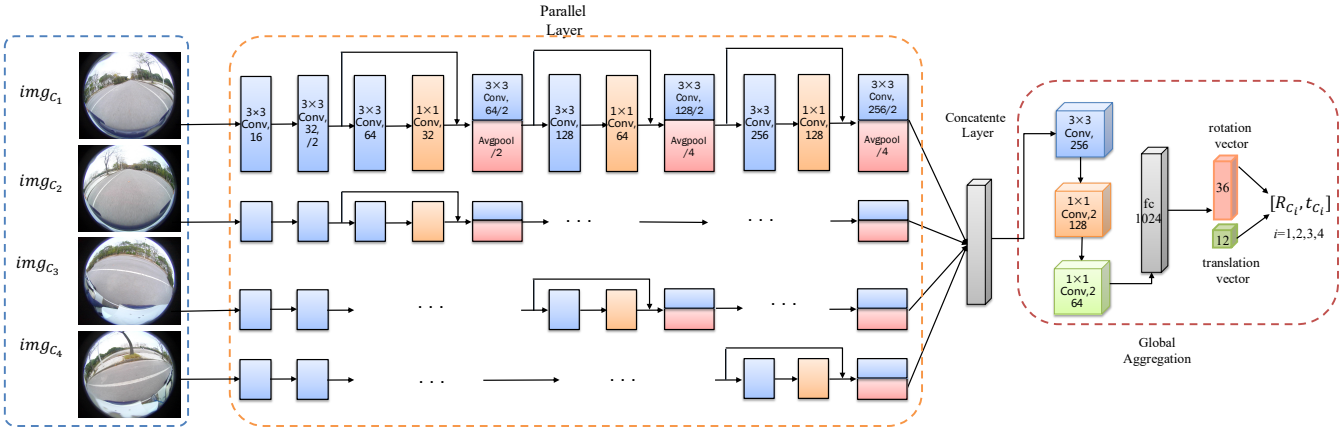


Fig. 1. The network architecture of WESNet. As illustrated in the figure, the parallel layer will extract deep-level features from four input images independently, and then the concatenate layer will fuse the features from different images. Finally, the global aggregation layer will map the concatenated features to a 48-dimensional vector which stands for the extrinsics of the SVS.

As shown in Fig. 1, the parallel layers independently extract deep-level features from each input fisheye image. Then the feature maps from multipath will be concatenated together so as to fuse the extracted features from different views. Finally, the global aggregation layer, which consists of both convolutional layers and fully connected layers, will map the aggregated features to the extrinsics. As mentioned in Eq. 1, these extrinsics are formulated in the transformation matrix form, and for each camera, nine rotation parameters and three translation ones need to be regressed. Thus, the WESNet will yield a 48-dimensional vector to estimate the extrinsics of the SVS.

B. Loss Function

The loss function of WESNet is defined as the composition of three loss terms, the geometric loss, the orthogonal one and the photometric one, which is given as,

$$Loss = Loss_{geo} + \alpha Loss_{ortho} + \beta Loss_{pho} \tag{8}$$

where $Loss_{geo}$ is the geometric loss, $Loss_{ortho}$ stands for the orthogonal one, and $Loss_{pho}$ refers to the photometric one. The geometric loss is actually the weakly supervised loss, which can promote the convergence of the network, while the photometric loss is to fine-tune the network so as to synthesize seamless surround-views. Besides, an orthogonal loss is also integrated to keep the internal constraints of the estimated rotation matrices. Next, we will introduce these three parts in detail.

Geometric loss. As we know, the performance of supervised learning strongly depends on the accuracy of labels, while in the task of extrinsics calibration, it is difficult to obtain accurate GTs. Therefore, we do not take the results of existing calibration solutions for fully supervised learning, but directly utilize the geometric relationships in the calibrated images as weak supervision information to construct the geometric loss guiding the optimization of the network, so as to avoid the negative impact the algorithm error of the reference schemes bringing to our network performance.

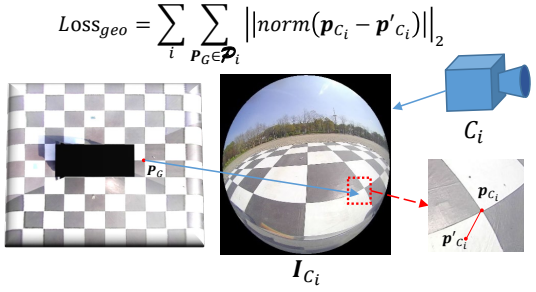


Fig. 2. Illustration of the geometric loss. It is constructed by calculating the pixel differences between the target corners and points projected from ground coordinate sysytem to pixel coordinate sysytem by Eq. 2.

We firstly collected the calibrated images over the calibration site filled with chessboard markers. With the known sizes of these markers, the 3D coordinates in the world coordinate system of each marker can be easily obtained. Based on the pair of 3D world coordinates of markers and manually labelled 2D pixel coordinates in calibrated images, we form the re-projection loss to geometrically regress the output extrinsics from images collected under the same extrinsics configurations. To help understand, the sketch of the geometric loss is illustrated in Fig. 2.

Given a selected corner on the calibration site, the relationship between its 3D coordinate $\boldsymbol{P}_G$ in the ground coordinate system $\boldsymbol{O}_G$ and its 2D pixel coordinate $\boldsymbol{p}_{C_i}$ on the undistorted fisheye image collected by camera C_i has been given in Eq. 2. Thus, with the yielded camera pose of WesNet, a corresponding projection $\boldsymbol{p}'_{C_i}$ can be generated and a loss term can be established, which is the square of the distance between $\boldsymbol{p}_{C_i}$ and $\boldsymbol{p}'_{C_i}$. By summing up the error terms of all corners, we obtain the overall geometric loss of WesNet, which is given as,

$$\begin{aligned} Loss_{geo} &= \sum_i \sum_{\boldsymbol{P}_G \in \boldsymbol{\mathcal{P}}_i} \left\| norm(\boldsymbol{p}_{C_i} - \boldsymbol{p}'_{C_i}) \right\|_2 \\ &= \sum_i \sum_{\boldsymbol{P}_G \in \boldsymbol{\mathcal{P}}_i} \left\| norm(\boldsymbol{p}_{C_i} - \frac{1}{Z_{C_i}} \boldsymbol{K}_{C_i} \boldsymbol{T}_{C_i} \boldsymbol{P}_G) \right\|_2 \end{aligned} \tag{9}$$

where i stands for the index of the camera in the SVS, ranging from one to four, $\mathcal{P}_i$ stands for the set of all selected corners on the calibration site that can be seen by camera C_i, the function $norm(*)$ means normalizing the re-projected pixel coordinate by dividing its coordinate on the corresponding axis with the width or the height of the image, respectively.

It's worth mentioning that the ground coordinate system should be determined manually. A common solution is to park the vehicle at the appropriate position over the calibration site to align the vehicle coordinate system and the ground one. From this viewpoint, except for labelling the training data in an weakly supervised manner, the geometric loss also offers an absolute reference information to guarantee the convergence of the network. Specifically, by introducing the geometric loss, WESNet can learn to determine a specific ground coordinate system, and regress poses of different cameras in a unified reference system.

Orthogonal loss. The rotation matrix consists of nine parameters but its DoF is only three, implying strong internal constraint. Since it's difficult to solve the constrained optimization problem, we choose a relatively soft solution, that is, introducing the orthogonal loss to keep the constraint satisfied along with the training process. Motivated by [41], the orthogonal loss is defined as,

$$\begin{aligned} Loss_{org} = &\sum_{i}\sum_{j=1}^{3}((R_{j1}^{Ci})^2+(R_{j2}^{Ci})^2+(R_{j1}^{Ci})^2-\rho)^2 \\ &+(R_{11}^{Ci}R_{21}^{Ci}+R_{12}^{Ci}R_{22}^{Ci}+R_{13}^{Ci}R_{23}^{Ci})^2 \\ &+(R_{11}^{Ci}R_{31}^{Ci}+R_{12}^{Ci}R_{32}^{Ci}+R_{13}^{Ci}R_{33}^{Ci})^2 \\ &+(R_{21}^{Ci}R_{31}^{Ci}+R_{22}^{Ci}R_{32}^{Ci}+R_{23}^{Ci}R_{33}^{Ci})^2 \end{aligned} \tag{10}$$

where R_{jk}^{Ci} is the element in the j-row and the k-th column of the rotation matrix $\boldsymbol{R}_{C_i}$ of C_i, and ρ tends to be one.

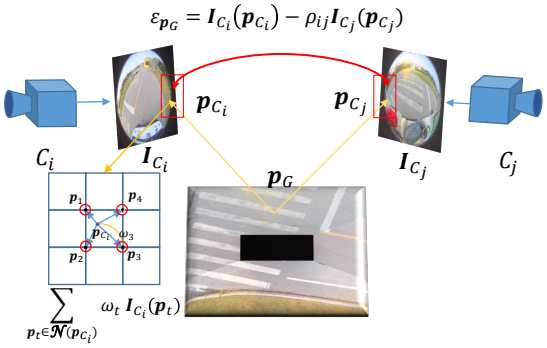


Fig. 3. Illustration of the photometric loss. It will be introduced in the second stage of the training approach to fine-tune the network, so as to synthesize seamless surround-views.

Photometric loss. Training only with the geometric loss, the network can converge to a barely satisfactory level. However, the results are still not accurate enough, and there will usually be obvious geometric misalignments in the synthesized surround-views. To further improve the performance of the network, we introduce an extra self-supervised photometric loss. The sketch of this photometric loss is illustrated in Fig. 3.

When the extrinsics are absolutely accurate, the grayscale values of the adjacent camreras' imaging of the same point tend to be consistent. Based on such an assumption, given a point $\boldsymbol{p}_G$ on the surround-view in the common-view region of camera C_i and C_j, we define the corresponding photometric loss term $\varepsilon_{\boldsymbol{p}_G}$ as,

$$\varepsilon_{\boldsymbol{p}_G} = \boldsymbol{I}_{GC_i}(\boldsymbol{p}_G) - \rho_{ij}\boldsymbol{I}_{GC_j}(\boldsymbol{p}_G) \tag{11}$$

where $\boldsymbol{I}_{GC_i}$ and $\boldsymbol{I}_{GC_j}$ are bird's-eye views of C_i and C_j. With the preliminaries given by Eq. 6, Eq. 11 can also be reformulated as,

$$\begin{aligned} \varepsilon_{\boldsymbol{p}_G} =&\boldsymbol{I}_{C_i}(\frac{1}{Z_{C_i}}\boldsymbol{K}_{C_i}\boldsymbol{T}_{C_iG}\boldsymbol{K}_G^{-1}\boldsymbol{p}_G) \\ &-\rho_{ij}\boldsymbol{I}_{C_j}(\frac{1}{Z_{C_j}}\boldsymbol{K}_{C_j}\boldsymbol{T}_{C_jG}\boldsymbol{K}_G^{-1}\boldsymbol{p}_G) \end{aligned} \tag{12}$$

where ρ_{ij} is an exposure factor to weaken the negative impact brought by the discrepancy on lighting conditions and environmental reflections between different cameras. Actually, as discussed in [42], for an image taken of a physical object, except for the properties of the object itself, the imaging pixel values will also be determined by the exposure time, the vignette and the non-linear response function of the camera. Among them, the exposure time is the most important factor according to our experience. Thus, we model the exposure factor ρ_{ij} as,

$$\rho_{ij} = \frac{t_i}{t_j} \tag{13}$$

where t_i is the corresponding exposure time of $\boldsymbol{I}_{C_i}$ and t_j is that of $\boldsymbol{I}_{C_j}$. Even though the exposure time can't be obtained directly in general, the factor ρ_{ij} can be fitted as,

$$\rho_{ij} = \frac{\sum_{\boldsymbol{p_G}\in\boldsymbol{\mathcal{O}}_{ij}}\boldsymbol{I}_{GCi}(\boldsymbol{p_G})}{\sum_{\boldsymbol{p_G}\in\boldsymbol{\mathcal{O}}_{ij}}\boldsymbol{I}_{GCj}(\boldsymbol{p_G})} \tag{14}$$

where $\boldsymbol{\mathcal{O}}_{ij}$ is the set of all pixels in the common-view region of C_i and C_j on bird's-eye-view images.

To improve the robustness to outliers, we adopted the L1 loss. In this way, by summing up the L1-norm of photometric loss terms $\varepsilon_{\boldsymbol{p}_G}$ corresponding to all qualified points $\boldsymbol{p}_G$, the overall photometric loss can be obtained as,

$$Loss_{pho} = \sum_{(i,j)\in\boldsymbol{\mathcal{A}}_{ij}}\sum_{\boldsymbol{p_G}\in\boldsymbol{\mathcal{N}}_{ij}}|\varepsilon_{\boldsymbol{p}_G}| \tag{15}$$

where $\boldsymbol{\mathcal{A}}_{ij}$ is the set of all adjacent cameras' indices, $\boldsymbol{\mathcal{N}}_{ij}$ is the set of all selected qualified pixels in the common-view region of C_i and C_j. For more details on the pixel selection, please refer to Sect. IV-C.

C. Implementation Details

Pixel selection. In the pixel selection approach, each qualified point $\boldsymbol{p}_G$'s intensity gradient modulus $G(\boldsymbol{p}_G)$ should be large enough so as to keep the stability of the feature it can offer. Specifically,

$$G(\boldsymbol{p}_G) \geq 2(G_{mean} + \sigma_g) \tag{16}$$

where G_{mean} is the mean gradient modulus over $\mathcal{O}_{ij}$, and σ_g is the associated standard deviation.

Besides, as discussed in Sect. III, for any point $\boldsymbol{p}_G$ on the surround-view, it is assumed to be on the ground. However, some objects with non-negligible heights, such as pedestrains, lawns or curbs, may appear in the surround-view and break such a preliminary. For ease of representation, we call these objects as "mismatched objects" and the corresponding pixels as "mismatched pixels". Constructing the photometric loss with mismatched pixels may do harm to the final accuracy since such pixels do not follow the imaging model of the SVS. Thus, such pixels should be eliminated in the pixel selection approach. Motivated by [21], we adopted a color based strategy. Specifically, for any qualified point $\boldsymbol{p}_G$, the color discrepancy between $\boldsymbol{I}_{GC_i}(\boldsymbol{p}_G)$ and $\boldsymbol{I}_{GC_j}(\boldsymbol{p}_G)$ is supposed to be unobvious. Defining $\boldsymbol{I}^c_{GC_i}$ and $\boldsymbol{I}^c_{GC_j}$ be the map of $\boldsymbol{I}_{GC_i}$ and $\boldsymbol{I}_{GC_j}$ of channel c, respectively, we measure the color discrepancy with the standard deviation of $\boldsymbol{p}_G$'s color ratios in different channels,

$$D_{color}\left(\boldsymbol{p}_G\right)=\sqrt{\frac{\sum_{c=1}^{n_c}\left(r_c\left(\boldsymbol{p}_G\right)-r_\mu\left(\boldsymbol{p}_G\right)\right)^2}{n_c}} \tag{17}$$

where n_c is the number of channels (normally 3) and r_μ is the average of all $\boldsymbol{p}$'s color ratios. The color ratio $r_c\left(\boldsymbol{p}_G\right)$ is defined as,

$$r_c\left(\boldsymbol{p}_G\right)=\frac{\boldsymbol{I}^c_{GC_i}\left(\boldsymbol{p}_G\right)}{\boldsymbol{I}^c_{GC_j}\left(\boldsymbol{p}_G\right)} \tag{18}$$

For any qualified $\boldsymbol{p}_G$, it must satisfy,

$$D_{color}\left(\boldsymbol{p}_G\right)<D_{mean}-2\sigma_d \tag{19}$$

where D_{mean} is the average color discrepancy of all the points in $\boldsymbol{\mathcal{O}}_{ij}$ and σ_d is the associated standard deviation.

Derivatives of the photometric loss. During the back propagation of the network, the derivation of the aforementioned three types of losses to the yielded extrinsics are indispensable. Different from the geometric loss and the orthogonal one, there is no perfect analytical solution of the derivatives of the photometric loss, thus some approximations are necessary. In this paper, we mainly discuss the specific derivation scheme of the photometric loss. Take the derivative δ of the photometric loss term $\varepsilon_{\boldsymbol{p}_G}$ to the pose matrix $\boldsymbol{T}_{C_iG}$ as an example. The corresponding derivative δ can be decomposed into multiple simpler parts via the chain rule,

$$\begin{aligned}\delta=&\frac{\partial\varepsilon_{\boldsymbol{p}_G}}{\partial\boldsymbol{I}_{C_i}}\cdot\frac{\partial\boldsymbol{I}_{C_i}}{\partial\boldsymbol{p}^T_{C_i}}\cdot\frac{\partial\boldsymbol{p}_{C_i}}{\partial\boldsymbol{P}^T_{C_i}}\cdot\frac{\partial\boldsymbol{P}_{C_i}}{\partial\boldsymbol{T}_{C_iG}}\\=&\frac{\partial\boldsymbol{I}_{C_i}}{\partial\boldsymbol{p}^T_{C_i}}\cdot\frac{\partial\boldsymbol{p}_{C_i}}{\partial\boldsymbol{P}^T_{C_i}}\cdot\frac{\partial\boldsymbol{P}_{C_i}}{\partial\boldsymbol{T}_{C_iG}}\end{aligned} \tag{20}$$

where $\boldsymbol{p}_{C_i}$ is the pixel coordinate and $\boldsymbol{P}_{C_i}$ is the corresponding coordinate in the camera system of C_i. The analytical solutions of the latter two terms can be derived easily, but not the first one. In [43], Irani *et al.* offered a general solution, that is utilizing the gradient of image intensities at $\boldsymbol{p}_{C_i}$ for approximation. In our implementations, to calculate the derivatives under the framework of neural network efficiently, each time the forward propagation finished, we linearize the photometric loss term $\varepsilon_{\boldsymbol{p}_G}$ at the current projection $\hat{\boldsymbol{p}}_{C_i}$ with the differentiable bilinear sampling and $\boldsymbol{I}_{C_i}(\boldsymbol{p}_{C_i})$ is reformulated as,

$$\boldsymbol{I}_{C_i}(\boldsymbol{p}_{C_i})=\sum_{\boldsymbol{p}_t\in\boldsymbol{\mathcal{N}}(\hat{\boldsymbol{p}}_{C_i})}\omega_t\cdot\boldsymbol{I}_{C_i}(\boldsymbol{p}_t) \tag{21}$$

where $\boldsymbol{\mathcal{N}}(\hat{\boldsymbol{p}}_{C_i})$ is the set of all neighboring points near $\hat{\boldsymbol{p}}_{C_i}$, and ω_t is linearly proportional to the spatial proximity between $\hat{\boldsymbol{p}}_{C_i}$ and $\boldsymbol{p}_t$. Through the reformulation, the undifferentiable term $\boldsymbol{I}_{C_i}(\boldsymbol{p}_{C_i})$ are converted to a differentiable one, and the back propagation of WESNet can be efficiently conducted under the auto-differential framework.

V. Surround-view Dataset

Since there's no existing large-scale surround-view dataset containing original fisheye images, we collected our own dataset by an open-topped electric car equipped with four cameras mounted around, which mainly consists of two parts, calibrated images and natural ones. The calibrated images were collected over the calibration site to provide weak supervision information, while the natural ones, which act as training and testing sets, were taken from natural scenes. The original resolutions of all collected images are 1920 × 1080 (1080p).

A. Calibrated Images

As aforementioned, the calibrated images were taken over a calibration site. The calibration site is located on a flat field with 10 × 10 chessboard grids printed on it, and the size of each grid is 1m × 1m, as shown in Fig. 4. We parked the vehicle to a designated position where the midpoint of rear axle of wheels is five meters horizontally and six meters vertically from the upper left corner of the calibration site. Defining this midpoint as the origin, the world coordinate system is established and the 3D world coordinate of each chessboard corner can be easily known.

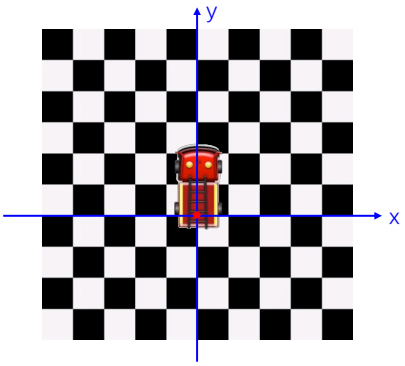


Fig. 4. Illustration of the calibration site and how to establish world coordinate system. Over the site, calibrated images, which offer weakly supervised information, will be collected.

B. Natural Images

We simulated the driving in reality and collected a large number of images in the natural environment on campus, namely natural images for short. The data we collected covers

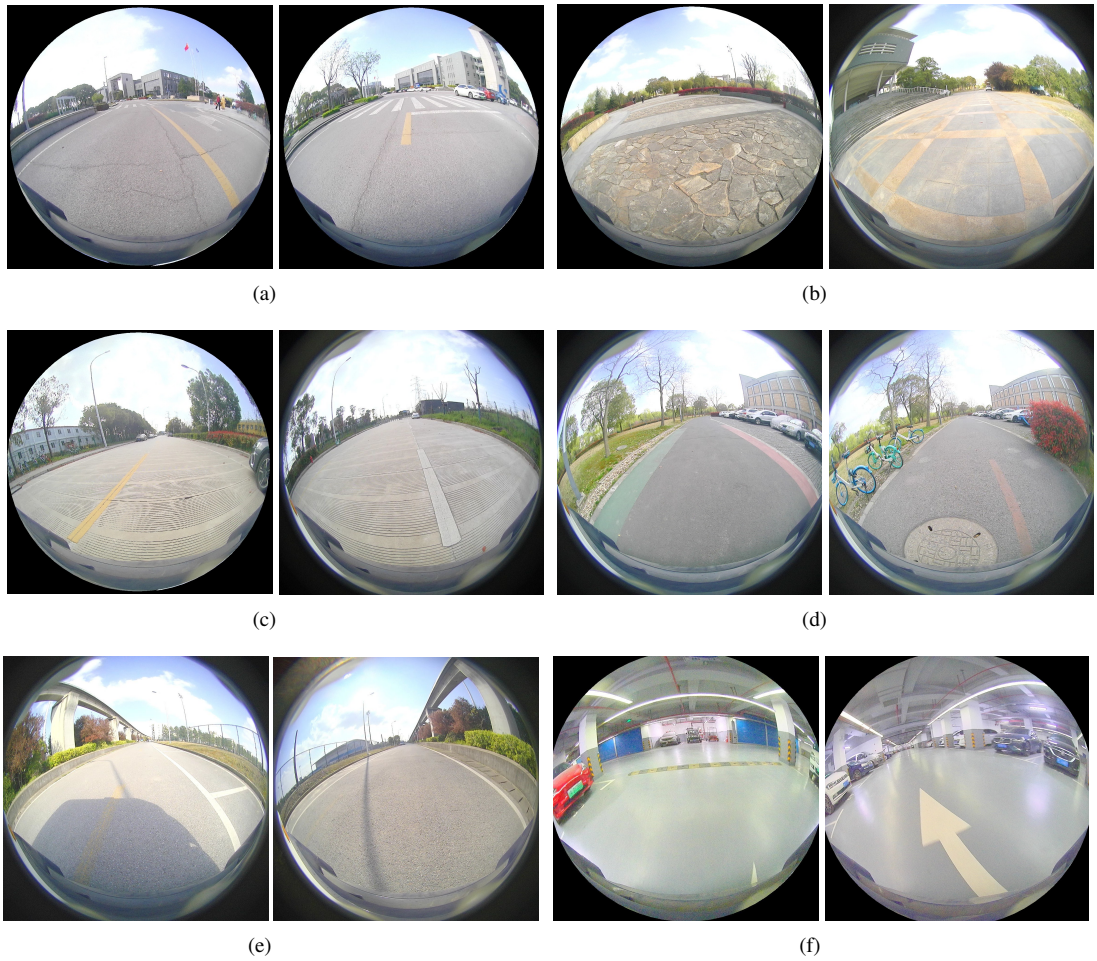


Fig. 5. Typical samples of natural images under different environments. From (a)∼(f), the shown images are all captured by the front camera and correspond to group V0∼V5 in Table I, respectively.

a variety of common environments, including the ground with different lane-lines and tiles, the cement road, the narrow path, and the road exposed to strong sunlight. In addition, due to the obvious difference between the underground environment and the above outdoor scenes, we also collected data from three different scales of underground garages. Some typical examples of different categories are shown in Fig. 5 while the specific quantities of images are in Table I.

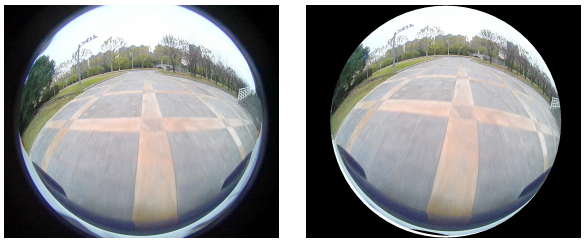
TABLE I
QUANTITIES OF NATURAL IMAGES IN DIFFERENT ENVIRONMENT

index	environment	specific type	numbers	total
V0	lane-line	——	6,873	6,873
V1	tile	square rectangle granite	1,599 2,001 312	3,912
V2	cement	——	526	526
V3	narrow	——	1,256	1,256
V4	sunlight	——	1,692	1,692
V5	garage	small medium large	736 1,486 2,597	4,819
***total*:19,078**				

C. Data Pre-processing

Preparing for the training, the collected data needs to be pre-processed, including labelling and data augmentation. For the labelling aspect, we recorded the pixel coordinates and the corresponding 3D world coordinates of the manually selected corners in the calibrated images as weakly supervised labels. For each frame, about 20∼30 corners were chosen. It is worth mentioning that this is the only necessary manual operation in our scheme. For the data augmentation aspect, in reality, the extrinsics of the SVS equipped by different vehicles usually vary, thus the collected data should cover extrinsics as widely as possible. However, it is quite time- and labor-consuming to expand the diversity of extrinsics of the dataset by manually adjusting the camera pose repeatedly and then collect data under different extrinsics' configurations for many times. Therefore, in our practice, the extrinsics of cameras were always fixed during the collection, while the homography transformation was applied to improve the richness of the extrinsics of collected data. Concretely, we applied the homography transformation to the normalized planes of undistorted images and then distorted them again to synthesize fisheye images corresponding to new extrinsics. Fig. 6 shows an example of the fisheye image before and after the augmentation. In this augmentation, the rotation

disturbance is within ± 0.2 rad (± 17 degrees), with an interval of 0.05 rad while the translation disturbance is within ± 0.2m, with an interval of 0.05m. Finally, the extrinsics' settings are augmented into nine different types, and the dataset containing 170,000 groups of images (each group consists of four fisheye images collected synchronously) is established. The 10% of the data is splited to the testing set (69,000 groups), and the rest forms the training set (155,000 groups). It is worth mentioning that we can transform the labels corresponding to the original calibrated images to generate those of augmented data instead of manually marking them, so that the data augmentation is fully automatic.



(a) (b)

Fig. 6. An example of synthesizing fisheye images based on our proposed data augmentation pipeline. (a) is the original fisheye image, and (b) is the augmented result, which can represent the fisheye image under a new extrinsics' configuration.

VI. Experimental Results

A. Experiment Setup

The training and evaluation code of our method are implemented using PyTorch [44]. Before training, all the input images were resized to 512×512. During training, we initialized the weights of all neural network layers randomly and used Adam optimizer [45] with 10^{-4} as the initial learning rate. Finally, we trained our network on Nvidia Titan Xp with batch size of 16 for 20 epochs.

B. Qualitative Experiment

TABLE II
Qualitative comparison with related methods in method features

method	prior	degree of automation	calibration target
Liu *et al.* [10]	×	manual calibration	tiles
Hedi *et al.* [11]	×	manual calibration	calibration site
Zhang *et al.* [12]	×	manual calibration	calibration chart
Shao *et al.* [13]	×	manual calibration	chessboard
Gao *et al.* [14]	×	manual calibration	calibration site
Zhang *et al.* [15]	×	manual calibration	chessboard
Zhao *et al.* [16]	×	self-calibration	ground lane
Choi *et al.* [17]	×	self-calibration	ground lane
Heng *et al.* [18]	×	self-calibration	SLAM system
Heng *et al.* [19]	×	self-calibration	SLAM system
Liu *et al.* [20]	√	self-calibration	natural images
OECS [21]	√	self-calibration	natural images
Ours	×	self-calibration	natural images

Traits of the methods. As we have reviewed in Sect. II, there are several studies in the literature that are relevant to our work in this paper. In order to understand the different characteristics of these methods more clearly, in Table II we compare them in three aspects: 1) Does it require the prior information of extrinsics? 2) Does it belong to manual calibration schemes or self-calibration ones? and 3) What kind of calibration targets does it rely on? It can be seen that Liu *et al.*'s method [20], OECS [21] and our scheme, WESNet, can yield the extrinsics of the SVS just from natural images. Among them, our scheme does not need any prior information of the extrinsics, implying a wider application scope. It's worth mentioning that Heng *et al.*'s schemes [18] and [19] also rely on natural images. However, as mentioned in Sect. II, since quantities of frames are required for their SLAM systems to converge, these two schemes are quite cubersome. For comparison, our scheme takes only a single group of frames collected by the SVS synchronously as the input, which corroborates the lightweight of WESNet.

Typical samples of synthesized surround-views. In order to qualitatively demonstrate the superiority of WESNet, we selected some typical samples from the collected data, and showed the surround-views sythesized with the yielded extrinsics of our scheme and two representative competitors, including the manual calibration scheme [13] and the self-calibration one, OECS [21], in Fig. 7. From Fig. 7, it can be clearly seen that there are often geometric misalignments of different severity in the synthesized surround-views corresponding to Shao *et al.*'s scheme [13]. This is because the offline calibration is only conducted over the calibration site, and is not adaptive to some naturally occurring interference factors, such as change of the tire pressure or the vehicle load. Besides, only a limited number of corners are utilized during the calibration, which causes sensitivity to labelling error of selected corners. And our method also suffers from the similar problem without the assistance of the photometric loss. In contrast, OECS [21] works well on the ideal road surface with clear texture, such as Fig.7 (b). However, since it only uses a single frame and strongly rely on the imaging hypothesis of the SVSs, its robustness is still unsatisfactory. When the road surface is not flat (as shown in Fig. 7 (f)) or there are obvious objects with non-negligible height (as shown in Fig. 7 (j) and (n)) in the surround-view, it may underperform. For our scheme, as aforementioned, since it can learn richer, deeper and more general features from a large number of data thanks to the superiority of the learning mechanism, in most cases, it performs much better than those two competitors. Fig. 7 also supports our claim.

C. Quantitative Experiment

Metrics. To measure the accuracy of the regressed extrinsics of compared methods, we mainly utilized two metrics, the re-projection error and the photometric error. The re-projection error refers to the distance between the projection of the 3D corners and the pixel coordinates of the corresponding 2D points in calibrated images. The photometric error is the grayscale difference between the corresponding keypoints, which lay in the common-view regions of the SVS, in the images captured by adjacent cameras. As the accuracy of

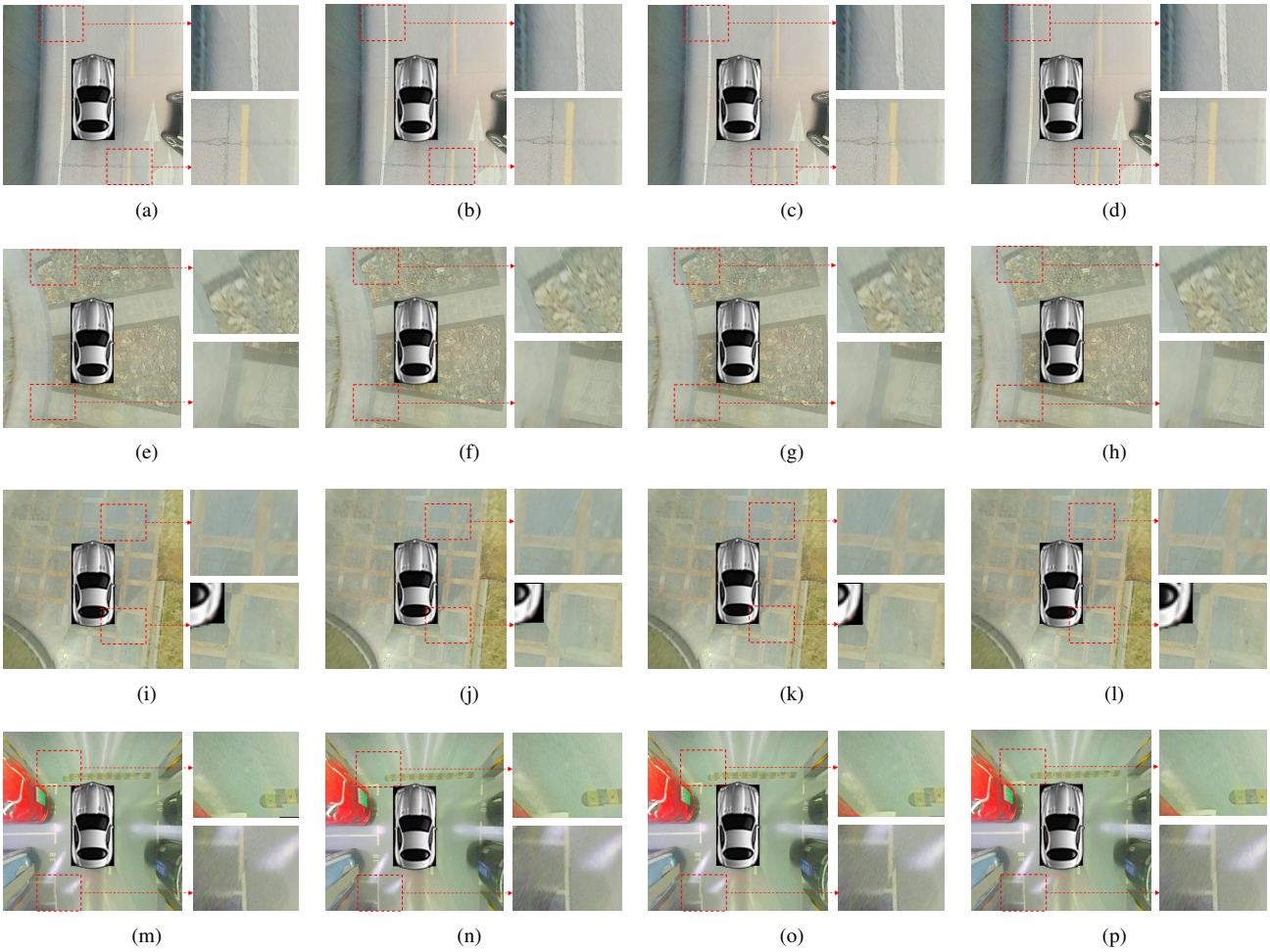


Fig. 7. Comparisons of the surround-views synthesized with yielded extrinsics of WESNet and two representative competitors. Besides, the WESNet trained without the photometric loss was also evaluated. In each row, from left to right the surround-views correspond to the result of Shao *et al.*'s scheme [13], OECS [21] and WESNet trained without and with the photometric loss, respectively.

offline calibration methods is generally satisfactory now, we take the re-projection and photometric errors over surround-views synthesized with camera poses offline calibrated by Shao *et al.*'s scheme [13] as the baseline (in short it can be called as "offline baseline"), and also offer the relative values of both errors to show the effectiveness of all competitors more intuitively. If the relative error is negative, it implies that the method is satisfactory. It's worth mentioning that since the calibrated images can just offer inaccurate supervision information with noise, the re-projection error is only for reference, while the photometric error is a more valuable metric.

Accuracy and generalization. Over each group of data we collected from different environments, we tested the compared methods with the metrics aforementioned, and summarized the experimental results in Table III. In addition, in order to more intuitively show the comprehensive performance, we weighted average the evaluation results of each metric according to the volume of group V0 $\sim$ V5, which is also shown in Table III. From the experimental results, it can be seen that our scheme without integrating the photometric loss performs best in the re-projection error aspect. However, its performance in the photometric aspect is quite unsatisfactory, implying that obvious over-fitting problem occurred. In terms of the photometric error, compared with all counterparts, our method shows an overwhelming performance over all datas except for group V2 after fusing the photometric information, which also reflects the excellent extrinsics estimation performance and the generalization of our method.

Robustness to intrinsic disturbance. To evaluate the robustness of our scheme to the accuracy of the intrinsics, we first introduced varying degrees of disturbance to the offline calibrated intrinsics of each camera in the SVS. The disturbance can be represented as an intrinsics' disturbance factor d and we added it to the focal length of the camera. Accordingly, the disturbed intrinsic matrix $K_{C_i}^d$ of camera C_i can be expressed as,

$$K_{C_i}^d = \begin{bmatrix} f_x + d & 0 & c_x \\ 0 & f_y + d & c_y \\ 0 & 0 & 1 \end{bmatrix} \tag{22}$$

Then under different d's settings and environment conditions we ran our scheme and recorded the corresponding relative re-projection and photometric errors. The relationship between the relative photometric errors and the settings of d is shown in Fig. 8. The figure illustrates that, as long as d is lower than

TABLE III
QUANTITIES OF GROUND IMAGES IN DIFFERENT ENVIRONMENT

environment	method	re-projection loss	relative reprojection loss	photometric loss	relative photometric loss
V0	Shao *et al*. [13]	8.2	0	29.5	0
	OECS [21]	4.5	-3.7	15.7	-13.8
	Ours$_{nopho}$	3.2	-5.0	16.6	-12.9
	Ours	3.4	-4.8	**15.4**	**-14.1**
V1	Shao *et al*. [13]	8.2	0	30.6	0
	OECS [21]	6.6	-1.6	22.0	-8.6
	Ours$_{nopho}$	3.1	-5.1	23.5	-7.1
	Ours	3.3	-4.9	**21.4**	**-9.2**
V2	Shao *et al*. [13]	8.2	0	31.4	0
	OECS [21]	5.6	-2.6	**19.6**	**-11.8**
	Ours$_{nopho}$	3.6	-4.6	21.8	-9.6
	Ours	4.9	4.6	20.3	-11.1
V3	Shao *et al*. [13]	8.2	0	33.2	0
	OECS [21]	5.5	-2.7	27.6	-3.1
	Ours$_{nopho}$	4.5	-3.7	30.7	-2.5
	Ours	5.1	-3.1	**27.2**	**-6.0**
V4	Shao *et al*. [13]	8.2	0	40.2	0
	OECS [21]	3.9	-4.3	27.1	-13.1
	Ours$_{nopho}$	3.5	-4.7	34.3	-5.9
	Ours	3.8	-4.4	**26.2**	**-14.0**
V5	Shao *et al*. [13]	8.2	0	26.5	0
	OECS [21]	4.2	-4.0	17.6	-8.9
	Ours$_{nopho}$	3.4	-4.8	21.8	-4.7
	Ours	4.1	-4.1	**16.9**	**-9.6**
weighted average	Shao *et al*. [13]	8.2	0	34.2	0
	OECS [21]	5.0	-3.2	20.0	-14.2
	Ours$_{nopho}$	3.4	-4.3	22.3	-11.9
	Ours	3.8	-4.4	**19.1**	**-15.1**

8 pixels, the camera poses after correction with our scheme are always more accurate than the offline calibrated results. Based on our experience, the camera's focal length variation caused by the natural collision or bumps won't exceed 5 pixels in general. Therefore, it can be seen that our algorithm is robust to the variations of intrinsics.

Time Cost Analysis. Following a weakly supervised framework, the repetitive workload in the calibration process was greatly reduced, and WESNet shows the speed performance far exceeding that of other competitors. To support our claim, we summarized the time cost to complete the whole calibration process of compared methods, and the results are given in Table IV. From this table, it can be seen that as a representative manual solution, Shao *et al.*'s method [13] takes about half an hour to finish the task, which confirms our claim that manual schemes are usually cumbersome. OECS [21] can effectively correct imprecise extrinsics of the SVS in an online manner, but about two seconds are still required. Compared with them, as a lightweight CNN, WesNet can regress the extrinsics end-to-end with a total time cost of about 12.7ms, implying a real-time speed performance.

TABLE IV
TIME COST FOR COMPARED METHODS

method	Shao *et al*. [13]	OECS [21]	Ours
time	About half an hour	2s	12.7ms

D. Ablation study of the photometric loss

In order to verify the efficacy of our proposed photometric loss, we conducted ablation experiments by training the network with and without it and compared the extrinsics' estimation effects of WESNet under both settings. The comparison mainly consists of two aspects, the quantitative one and the qualitative one, which are shown in Fig. 5 and Table. III, respectively. For the qualitative aspect, the accuracy of competitors mainly manifests in the qualities of synthesized surround-views, and for the quantitative aspect, four metrics mentioned in Sect. VI-C are adopted. It can be seen that without the photometric supervision, our scheme can show the comparable performance to the offline calibrated scheme, but it's still inferior to OECS [21]. And by introducing such a loss term, the accuracy performance of the network can be effectively enhanced and overperforms all counterparts. Thus, the photometric loss we proposed is corroborated to be an essential and significant mechanism for our network.

E. Failure Case Analysis

By observing and analyzing the experimental results, we found that the textures of the ground have an obvious influence on our network performance. On the one hand, when the ground texture is weak, the information contained in the image is relatively scarce. In this way, the impact of noise will be more notable, and WESNet will not be able to extract high-quality features in deep level, resulting in poor performance as shown in Fig .9 (a). On the other hand, when the ground is filled with repetitive fine-grained textures, our network may

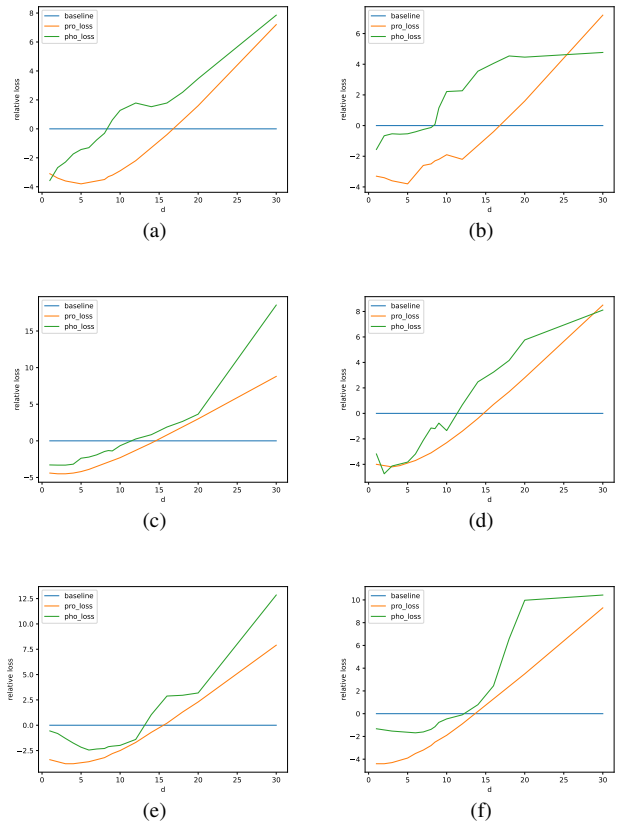


Fig. 8. The corresponding relative re-projection and photometric errors under different disturbance factor d's settings. The relationship between the relative re-projection errors and different settings of d is shown as the orange curve while that of relative photometric errors is shown as the green curve. The blue line is the offline baseline.

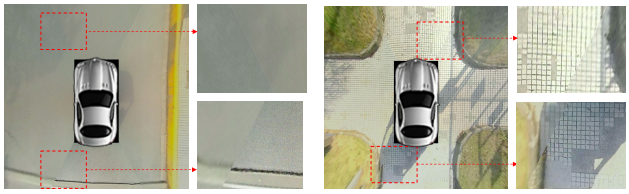


Fig. 9. Examples of failure cases. The synthesized surround-views are shown on the left while enlarged local regions are on the right.

also fail as shown in Fig .9 (b). This is because under our training framework, especially in the second stage with photometric loss, it tends to align the adjacent views in the SVS. In such a failure case, the alignment may be "mismatched", implying the failure of the extrinsics estimation. Thus, it should be emphasized that in order to make our method work successfully, the vehicle needs to be parked on flat ground with clearly observable and coarse grained textures.

VII. Conclusion and Future Work

In this paper, we studied a practical problem, extrinsics self-calibration for the surround-view system, emerging from the field of ADAS, and proposed a solution. Following a weakly supervised framework, we proposed a novel learning-based solution, namely WESNet. Taking the original fisheye images captured by cameras in the SVS as the input, it can yield extrinsics end-to-end. During the training approach, we first optimize the network fully based on weakly supervised geometric loss for fast convergence, and then the self-supervised photometric loss will be introduced to further fine-tune the network. With the two-stage training, seamless surround-views can be synthesized with the yielded extrinsics. An outstanding feature of WESNet is that, since the only required input is a single group of natrual fisheye images, and the forward propagation of the network can be completed in the milliseconds of time consumption, it does not require additional apparatuses or calibration sites and can be easily applied in the online manner. As long as the vehicle is driving on a normal flat road with relatively rich textures, our scheme will work. Besides, to facilitate the study of the extrinsics calibration or other surround-view based computer vision tasks, a surround-view dataset containing 19,078 groups of surround-views and the corresponding original fisheye images in high resolution was also collected where common environments are covered. As far as we know, currently this is the largest surround-view dataset publicly available which contains original fisheye images. Experimental results on it show that WESNet can promptly estimate the SVSs' extrinsics and synthesize seamless surround-views. However, up to now, the performance of our approach in the environment having low texture or strong texture repeatability is still not satisfactory and thus we will continue to devote efforts in this area.

References

[1] L. Duan and F. Chen, "The Future of Advanced Driving Assistance System Development in China," in *Proceedings of IEEE International Conference on Vehicular Electronics and Safety*, 2011, pp. 238-243.

[2] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge, U.K.: Cambridge Univ. Press, 2003.

[3] R. Klette, A. Koschan, and K. Schluns, *Computer Vision: Three-dimensional Data from Images*. Singapore: Springer, 1998.

[4] L. Li, L. Zhang, X. Li, X. Liu, Y. Shen, and L. Xiong, "Vision-Based Parking-Slot Detection: A Benchmark and a Learning-Based Approach," in *Proceedings of IEEE International Conference on Multimedia and Expo*, 2017, pp. 649-654.

[5] L. Zhang, J. Huang, X. Li, and L. Xiong, "Vision-Based Parking-Slot Detection: A DCNN-Based Approach and a Large-Scale Benchmark Dataset," *IEEE Transactions on Image Processing*, vol. 27, no. 11, pp. 5350-5364, 2018.

[6] C. Lin and M. Wang, "A Vision Based Top-View Transformation Model for a Vehicle Parking Assistant," *Sensors*, vol. 12, no. 4, pp. 4431-4446, 2012.

[7] J. Xu, G. Chen, and M. Xie, "Vision-Guided Automatic Parking for Smart Car," in *Proceedings of IEEE Intelligent Vehicles Symposium*, 2000, pp. 725-730.

[8] M. Gressmann, G. Palm, and O. Löhlein, "Surround View Pedestrian Detection Using Heterogeneous Classifier Cascades," in *Proceedings of International IEEE Conference on Intelligent Transportation Systems*, 2011, pp. 1317-1324.

[9] C. Hou, H. Ai, and S. Lao, "Multiview Pedestrian Detection Based on Vector Boosting." in *Proceedings of Asian Conference on Computer Vision*, 2007, pp. 18-22.

[10] Y. Liu, K. Lin, and Y. Chen, "Bird's-Eye View Vision System for Vehicle Surrounding Monitoring," in *International Workshop on Robot Vision*, 2008, pp. 207–218.

[11] A. Hedi and S. Lonari, "A System For Vehicle Surround View," *IFAC Proceedings Volumes*, vol. 45, no. 22, pp. 120–125, 2012.

[12] B. Zhang, V. Appia, I. Pekkucuksen, Y. Liu, A. Batur, P. Shastry, S. Liu, S. Sivasankaran and K. Chitnis,"A Surround View Camera Solution for Embedded Systems," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2014, pp. 676-681.

[13] X. Shao, X. Liu, L. Zhang, S. Zhao, Y. Shen, and Y. Yang, "Revisit Surround-View Camera System Calibration," in *Proceedings of IEEE International Conference on Multimedia and Expo*, 2019, pp. 1486-1491.

[14] Y. Gao, C. Lin, Y. Zhao, X. Wang, S. Wei and Q. Huang, "3-D Surround View for Advanced Driver Assistance Systems," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 1, pp. 320-328, 2018.

[15] L. Zhang, J. Chen, D. Liu, Y. Shen and S. Zhao, "Seamless 3D Surround View with a Novel Burger Model," in *IEEE International Conference on Image Processing*, 2019, pp. 4150-4154.

[16] K. Zhao, U. Iurgel, M. Meuter, and J. Pauli, "An Automatic Online Camera Calibration System for Vehicular Applications," in *Proceedings of International IEEE Conference on Intelligent Transportation Systems*, 2014, pp. 1490-1492.

[17] K. Choi, H. Jung, and J. Suhr, "Automatic Calibration of an Around View Monitor System Exploiting Lane Markings," *Sensors*, vol. 18, no. 9, pp. 2956-2982, 2018.

[18] L. Heng, B. Li, and M. Pollefeys, "CamOdoCal: Automatic Intrinsic and Extrinsic Calibration of a Rig With Multiple Generic Cameras and Odometry," in *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2013, pp. 1793-1800.

[19] L. Heng, M. Bürki, G. Lee, P. Furgale, R. Siegwart, and M. Pollefeys, "Infrastructure-Based Calibration of a Multi-Camera Rig," in *Proceedings of IEEE International Conference on Robotics and Automation*, 2014, pp. 4912-4919.

[20] X. Liu, L. Zhang, Y. Shen, S. Zhang, and S. Zhao, "Online Camera Pose Optimization for the Surround-View System," in *Proceedings of ACM International Conference on Multimedia*, 2019, pp. 383-391.

[21] T. Zhang, L. Zhang, Y. Shen, Y. Ma, S. Zhao and Y. Zhou, "OECS: Towards Online Extrinsics Correction For The Surround-View System" in *Proceedings of IEEE International Conference on Multimedia and Expo*, 2020, pp. 1-6.

[22] Z. Zhou, "A Brief Introduction to Weakly Supervised Learning," *National Science Review*, vol. 5, no. 1, pp. 44-53, 2018.

[23] L. Jing, Y. Tian, "Self-Supervised Visual Feature Learning With Deep Neural Networks: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 9, no. 9, pp. 1-22, 2020.

[24] C. Zhou, D. Tan, F. Zhu and Z. Dong, "A Planar Homography Estimation Method for Camera Calibration," in *Proceedings of International IEEE Conference Symposium on Computational Intelligence in Robotics and Automation*, 2003, pp. 424-429.

[25] C. Harris and M. Stephens, "A Combined Corner and Edge Detector," in *Proceedings of Alvey Vision Conference*, 1988, pp. 147-151.

[26] M. Calonder, V. Lepetit, C. Strecha, P. Fua, "BRIEF: Binary Robust Independent Elementary Features," in *Proceedings of European Conference on Computer Vision*, 2010, pp. 778-792.

[27] R. Battiti, "First- and Second-Order Methods for Learning: Between Steepest Descent and Newton's Method," *Neural Computation*, vol. 4, no. 2, pp. 141-166, 1992.

[28] M. Lourakis, "Sparse Non-Linear Least Squares Optimization for Geometric Vision," in *Proceedings of European Conference on Computer Vision*, 2010, pp. 43-56.

[29] J. Moré, "The Levenberg-Marquardt Algorithm: Implementation and Theory," in "Numerical Analysis (Eds: G.A. Watson)", pp. 105-116, Berlin, Germany: Springer, 1978.

[30] D. Lowe, "Distinctive Image Features From Scale-Invariant Keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91-110, 2004.

[31] H. Bay, T. Tuytelaars, and L. Gool, "SURF: Speeded Up Robust Features," in *Proceedings of European Conference on Computer Vision*, 2006, pp. 404-417.

[32] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An Efficient Alternative to SIFT or SURF," in *Proceedings of IEEE International Conference on Computer Vision*, 2011, pp. 2564-2571.

[33] S. Workman, C. Greenwell, M. Zhai, R. Baltenberger, and N. Jacobs, "Deepfocal: A Method for Direct Focal Length Estimations," in *Proceedings of IEEE International Conference on Image Processing*, 2015, pp. 1369-1373.

[34] M. Giering, V. Venugopalan, and K. Reddy, "Multi-Modal Sensor Registration for Vehicle Perception via Deep Neural Networks," in *Proceedings of High Performance Extreme Computing Conference*, 2015, pp. 1-6.

[35] N. Schneider, F. Piewak, C. Stiller, and U. Franke, "Regnet: Multi-Modal Sensor Registration Using Deep Neural Networks," in *Proceedings of IEEE Intelligent Vehicles Symposium*, 2017, pp. 1803-1810.

[36] G. Iyer, R. Ram, J. Murthy and K. Krishna, "CalibNet: Geometrically Supervised Extrinsic Calibration Using 3D Spatial Transformer Networks," in *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2018, pp. 1110-1117.

[37] Z. Zhang, "Flexible Camera Calibration By Viewing a Plane From Unknown Orientations," in *Proceedings of International Conference on Computer Vision*, 1999, pp. 666-673.

[38] F. Du and M. Brady, "Self-Calibration of the Intrinsic Parameters of Cameras for Active Vision Systems," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 1993, pp. 477-482.

[39] H. Zhu, J. Yang, and Z. Liu, "Fisheye Camera Calibration with Two Pairs of Vanishing Points," in *Proceedings of International Conference on Information Technology and Software Engineering*, 2009, pp. 321-324.

[40] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770-778.

[41] M. Ahmed, E. Hemayed, A. Farag, "Neurocalibration: A Neural Network That Can Tell Camera Calibration Parameters," in *IEEE International Conference on Computer Vision*, 1999, pp. 463-468.

[42] J. Engel, V. Koltun and D. Cremers, "Direct Sparse Odometry," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 3, pp. 611-625, 2018.

[43] M. Irani and P. Anandan, "About Direct Methods," in *Proceedings of International Workshop on Vision Algorithms*, 1999, pp. 267-277.

[44] A. Paszke and S. Gross, *et al.*, "PyTorch: An Imperative Style, High-Performance Deep Learning Library," 2019, *arXiv:1912.01703.* [Online]. Available: https://arxiv.org/pdf/1912.01703.

[45] D. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," 2014, *arXiv:1412.6980.* [Online]. Available: http://arxiv.org/abs/1412.6980.