




Multi-cameras Calibration System Based Deep Learning Approach and Beyond: A Survey

Alaa Shakir Mahmood^{1*} , Bashar Taleb² , Adil Abdul-Wahab³ 

^{1,2,3} Computer Science Department, College of Science, University of Diyala, Diyala, 32001, Iraq
scicompms222310@uodiyala.edu.iq¹, alnuaimi_bashar@uodiyala.edu.iq², adil_alazzawi@uodiyala.edu.iq³

Abstract

The process of determining camera settings to deduce geometric attributes from recorded sequences is known as camera calibration. This process is essential in the fields of robotics and computer vision, encompassing both two-dimensional and three-dimensional applications. Traditional calibration methods, however, are time-consuming and require specific expertise. Recent endeavors have demonstrated that learning-based systems can replace the monotonous tasks associated with manual calibrations. Responses have been examined through a range of learning techniques, networks, geometric assumptions, and datasets. A thorough examination of camera calibration systems that rely on learning algorithms is offered in this paper, assessing their advantages and disadvantages. The primary categories of calibration presented are the regular pinhole camera model, distortion camera model, cross-sensor model, and cross-view model. These categories align with current research trends and have diverse applications. As there is no existing standard in this field, a large dataset of calibration has been created, which can be used as a public platform to assess the effectiveness of current methods. This collection consists of both artificially generated and genuine data, including images and videos obtained from various cameras in different locations. The difficulties faced will be analyzed, and alternative avenues for further research will be suggested in the next stage of this project. This survey represents the initial attempt to perform camera calibration using learning-based methods spanning a period of eight years. Our findings indicate that learning-based methods significantly reduce the time and expertise required for calibration while maintaining or improving accuracy compared to traditional methods. Specifically, our research demonstrates a calibration error reduction of up to 20% and speed improvements by a factor of three compared to traditional methods, as well as better adaptability to different camera types and environments.

Keywords: Camera Calibration, Deep Learning, Computer vision, Depth Estimation

Article history: Received: 24-5-2024, Accepted: 21-7-2024, Published:

This article is open-access under the CC BY 4.0 license (<http://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The area of camera calibration is considered crucial and essential in computer vision, with a significant research background dating back about 60 years [1]. The initial stage for numerous vision and robotics tasks involves the calibration of intrinsic (photo distortion and sensor parameters) and/or extrinsic (translation and rotation) parameters of cameras. This calibration process is applied to various fields such as computational photography, Multiview geometry, and 3D reconstruction. Regarding the task type, several methods have been

developed for calibrating different types of cameras, such as stereo cameras, pinhole cameras, fisheye lens cameras, LiDAR-camera systems, light field cameras, and event cameras.

The conventional techniques of camera calibration typically rely on manually designed assumptions and characteristics about the model. These approaches can be categorized into three broad types. One often-used method involves utilizing a well-known calibration target, such as a checkerboard, which is intentionally moved inside the three-dimensional image [2], [3]. The target is

* Corresponding author: scicompms222310@uodiyala.edu.iq

then recorded by the camera from several perspectives, and the corners of the checkerboard are identified to calculate the camera settings. However, this process necessitates laborious manual interventions and is incapable of achieving automatic calibration in real-world scenarios.

The second type of calibration, known as geometric-prior-based camera calibration, has been extensively researched to improve flexibility. This area of study has been explored in several publications, including in [4]. More precisely, geometric structures, such as lines and vanishing points, are utilized to represent the 3D-2D relationship in the picture. However, this approach strongly depends on controlled artificial scenes with abundant geometric priors, resulting in subpar performance when used in diverse situations. The third category is self-calibration, as described in references [5]. This technique utilizes a series of photos as inputs and applies Multiview geometry to estimate the camera parameters. However, it is

restricted by the limitations of the feature detectors, which can be affected by various lighting situations and textures. Due to the availability of numerous established methods for calibrating cameras in industrial or laboratory settings [6], this step is often overlooked in current advancements. Nevertheless, the process of calibrating individual and uncontrolled photos remains difficult, particularly when the images are sourced from websites and captured by unfamiliar camera models as displayed in Fig. 1. This problem serves as a motivation for academics to explore a novel paradigm.

Deep learning has recently provided fresh insights into camera calibration and its practical uses. Learning-based strategies consistently yield the best results on a wide range of activities, while also being more efficient. Various deep neural networks (DNNs) have been created, including generative adversarial networks (GANs), convolutional neural networks (CNNs), vision transformers (ViTs), and Point Net.

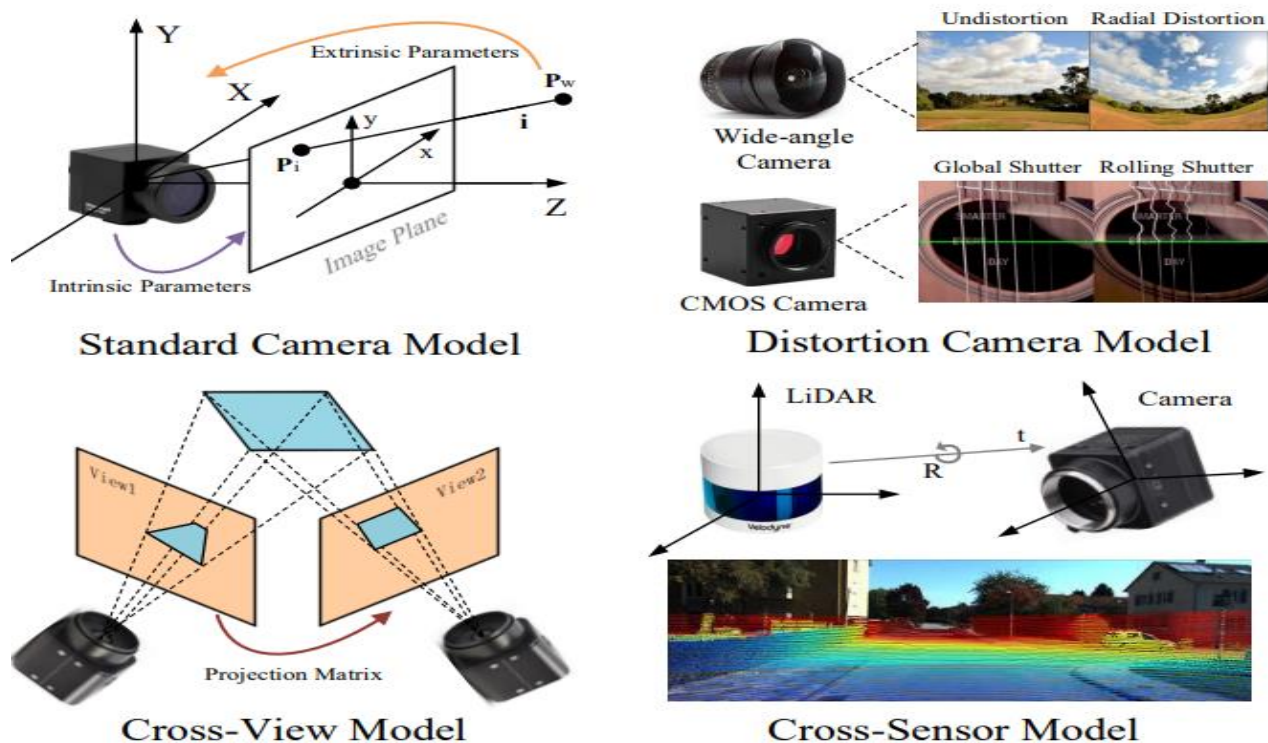


Fig. 1 Displays the often-used calibration models and expanded uses in camera calibration.

These DNNs exhibit superior representation capability in terms of high-level semantic features compared to manually designed features.

Furthermore, several learning strategies have been employed to improve the geometric perception of neural networks. Learning-based systems provide a

completely automated solution for camera calibration, eliminating the need for manual interventions or calibration objectives, and distinguishing them from previous methods. In addition, several techniques can perform calibration without relying on specific camera models or labels, demonstrating potential and significant practical uses.

Keeping pace with the continuous proliferation of learning-based camera calibration techniques has gotten progressively more difficult. Therefore, it is imperative to examine current literature and cultivate a community that is devoted to this area of study. Historically, specific surveys, such as [7], focused exclusively on a particular task or camera within the realm of camera calibration or only investigated one specific approach.

In the study of Salvi et al. [8], conducted a review of classic camera calibration methods, specifically focusing on the algorithms used. Hughes et al. [9], conducted a comprehensive analysis of the process of calibrating fisheye cameras using conventional methods. Although Fan et al. [10], examined both deep learning and standard approaches, their review primarily focuses on the calibration of wide-angle cameras. Furthermore, the limited number of evaluated learning-based approaches (about 10 publications) makes it challenging for readers to comprehend the progression of general calibration [10]. This study presents a thorough and detailed examination of the latest developments in camera calibration using learning-based methods. It encompasses an analysis of more than 100 studies. In addition, we explore potential avenues for further enhancements and analyze different categories of cameras and targets.

To streamline research of the future on various subjects, we classify the existing solutions based on calibration goals and practical uses. Aside from basic parameters like rotation, translation, and focal length, we also offer comprehensive evaluations for rectifying image distortion, determining and calibrating camera-LiDAR systems, cross-view mapping, and other uses. This trend is driven by advancements in camera technology and industry

demands for virtual reality, neural rendering, autonomous driving, and other related applications.

As far as we know, this is the initial investigation into the use of learning-based methods for camera calibration and its expanded applications. It offers the following distinct contributions.

- Our research primarily tracks current developments in deep learning-based camera calibration. thorough examination and conversation articles, network architecture, loss functions, datasets, evaluation metrics, learning algorithms, implementation platforms, etc. are among the many aspects that are provided.
- In addition to the calibration procedure, we thoroughly examine the traditional camera models and their expanded versions. Specifically, we outline the revised calibration targets in deep learning, as many conventional calibration objectives have been proven to be challenging for neural networks to acquire.
- We collected a dataset of 10,000 images and 500 videos from various cameras, including stereo, fisheye, pinhole, LiDAR, light field, and event cameras, captured in diverse environments such as indoor scenes, outdoor landscapes, urban areas, and rural settings. This dataset can be used to evaluate the generalization capability of current algorithms.
- We analyzed the unresolved difficulties of learning-based calibration and suggested potential avenues for future research to offer assistance in this domain.
- A publicly accessible repository is established to offer a systematic classification of all evaluated works and performance standards. Fig. 2 The categorization and organization of camera calibration using deep learning, based on its structure and hierarchy. Each category has a list of classical methods.

2. Preliminaries

The advent of deep learning has provided fresh insights into camera calibration, allowing for a completely automated calibration process. without any need for manual intervention. In this section, we provide an overview of two commonly used approaches in learning-based camera calibration: regression-based calibration and

reconstruction-based calibration. Next, this research subject examines the commonly employed learning methodologies. The comprehensive definitions for classical models and accompanying calibration objectives are presented in the supplemental material.

2.1 The Concept of Learning Paradigm

The researchers have created two basic paradigms for learning-based calibrated cameras and their applications, which are driven by various designs of the neural network. Calibration using regression analysis. When provided with an uncalibrated input, the regression-based calibration method initially utilizes stacked convolutional layers to extract high-level semantic characteristics. Next, the fully connected layers combine the semantic information and

generate a vector representing the estimated calibration aim.

The obtained parameters are utilized for carrying out the following activities, such as rectifying distortions, distorting images, localizing the camera, and so on. This paradigm is the most ancient and holds a prominent position in the process of learning-based camera calibration and its various uses. The paradigm has successfully achieved various objectives, such as Deep focal [11], for intrinsics, PoseNet for extrinsic [12], for radial distortion, URSCNN for rolling shutter distortion, DHN for homography matrix [11], for hybrid parameters, and RegNet for camera-LiDAR parameters.

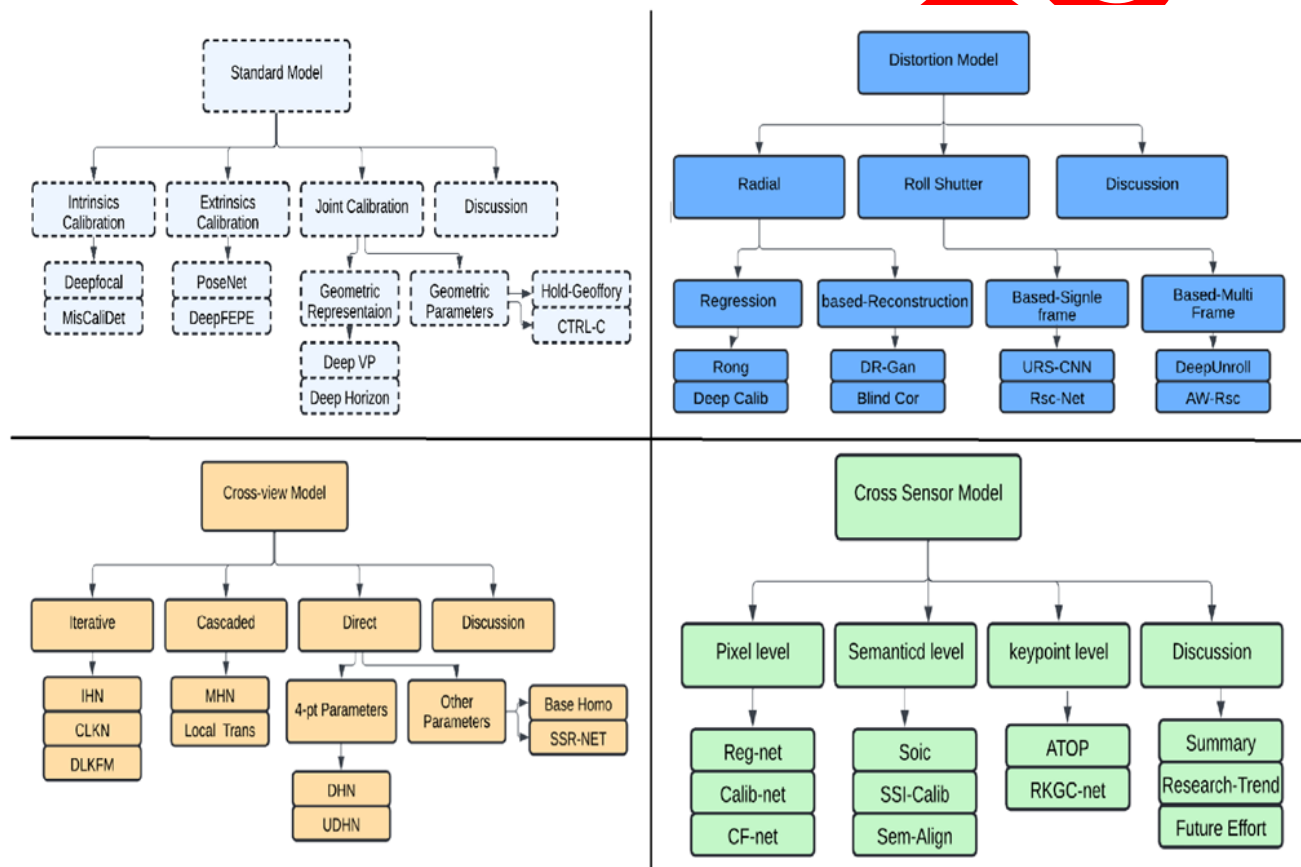


Fig. 2 The categorization and organization of camera calibration using deep learning, based on its structure and hierarchy, each category has a list of classical methods

Calibration using reconstruction-based methods. However, the reconstruction-based calibration paradigm eliminates the parameter regression and instead directly learns the pixel-level mapping function between the uncalibrated input and target. This approach is influenced by the concepts of conditional translation image-to-image [13], so dense visual perception. The reconstructed results are

subsequently computed for the pixel-wise loss using the ground truth. Regarding this matter, most calibration methods that rely on reconstruction develop their network architecture using a fully convolutional network, such as U-Net [14]. An encoder-decoder network is used, which includes skip links between the encoder and decoder. This network gradually extracts features from low-level to

high-level and efficiently combines multi-scale features. In the final convolutional layer, the acquired characteristics are combined into the desired channel, resulting in the reconstruction of the calibrated outcome at the individual pixel level. Unlike the regression-based approach, the reconstruction-based approach does not necessitate the inclusion of many camera parameters. In addition, the issue of imbalance loss can be resolved as the optimization process focuses just on the photometric loss of calibrated outcomes. Hence, the reconstruction-based paradigm allows for blind camera calibration without relying heavily on a specific camera model.

2.2 Acquisition Techniques

In the following, we examine the existing research on camera calibration using learning-based methods, focusing on various tactics employed in the literature. **Supervised learning:** Many camera calibration approaches that rely on learning train their networks using supervised learning strategies. These strategies range from classical methods to state-of-the-art methods [15]. Regarding the learning paradigm, this technique oversees the network using either the accurate parameters of the camera (regression-based paradigm) or paired data (reconstruction-based paradigm). Typically, they create the training dataset by combining various large-scale datasets, using random parameter sampling and camera model simulation. Several recent studies [16] have utilized a real-world setting to create their training dataset.

These studies have manually annotated the collected photos, which has contributed to **Semi-supervised learning:** Utilizing an annotated dataset in various contexts is a highly successful technique for training the network. Nevertheless, human annotation is susceptible to errors, which might result in inconsistent quality annotation or the inclusion of data. Therefore, it can be difficult to enhance performance by expanding the training dataset because of the intricate and expensive process of creating the dataset. In order to tackle this difficulty, SS-WPC [17], presents a semi-supervised approach for rectifying portraits taken with a wide-angle camera. The method utilizes a surrogate task (segmentation) and a semi-supervised approach that leverages direction and range consistency as well as

regression consistency to make use of both data (labeled and unlabeled).

Weakly-supervised learning: refers to a type of machine learning when the training data is only partially labeled or labelled at a coarse level of granularity. Despite making major advancements, the process of data labeling for camera calibration is well-known for being expensive, and it is difficult to produce completely accurate ground-truth labels. Consequently, it is frequently more advantageous to employ poor supervision in conjunction with machine learning techniques. Weakly supervised learning is the method of constructing prediction models with limited supervision. In the study of Zhu et al. [18], propose a weakly supervised calibration technique for single-view metrology in unconstrained environments, where only one image of a scene with objects of uncertain sizes is available. This work utilizes 2D object annotations from extensive databases, which often contain people and buildings. These items are valuable references for estimating the size of 3D things.

Unsupervised learning: also known as unsupervised machine learning, employs machine learning algorithms to analyze and categorize datasets that do not have labeled information. The UDHN [19], is the initial study that applies unsupervised learning to estimate the homography matrix of a paired picture in a cross-view camera model, without relying on projection labels. UDHN [19], surpasses prior supervised learning methods by minimizing pixel-wise intensity error without the need for ground truth data. The suggested unsupervised technique can achieve faster inference time while maintaining improved accuracy and robustness against fluctuations in light. Building upon this research, an increasing number of approaches are utilizing the unsupervised learning technique to estimate the homography. Some notable examples include CA-UDHN [20], Liu et al. [21], Base Homo [22], and Homo GAN [23]. In addition, Un FishCor [24], eliminates the need for parameters and develops a framework that is unsupervised to wide-angle cameras.

The supervised Self-learning term " supervised self-learning" originated in the field of robotics,

where automatically the training data is classified by leveraging the connections between different input sensor signals. Self-supervised learning utilizes the input data itself as the means of supervision, in contrast to supervised learning. Several self-supervised methods are introduced to acquire visual attributes from large quantities of unlabeled photographs or videos, eliminating the requirement for laborious and costly human annotations. The SSR-Net [22], introduces a self-supervised deep homography estimate network that eliminates the requirement for ground truth annotations and utilizes the invertibility limitations of homography. SSR-Net [25], employs the homography matrix representation instead of the commonly used 4-point parameterization in other techniques, to enforce invertibility restrictions. SIR proposes a novel self-supervised camera calibration pipeline for rectifying wide-angle photos.

The pipeline operates on the concept that the rectified outputs of distorted images captured with different lenses should be the same. The work of Fang et al. [26], the authors provides a method for self-calibrating various camera models using self-supervised depth and pose learning. This method allows for the calibration evaluation of camera parameters learned exclusively through self-supervision, using raw video data. Reinforcement Learning Reinforcement learning focuses on maximizing the overall advantages of a learning process, rather than minimizing them at each stage. As of now, DQN-RecNet [26], is the exclusive and pioneering research in camera calibration that utilizes reinforcement learning. This method utilizes a sophisticated form of reinforcement learning to address the task of fisheye picture rectification. It accomplishes this by employing a single Markov Process, which is a step-by-step calibration process. Here, the present fisheye image accurately depicts the condition of the surroundings.

3. Standard Model

The main objectives of intrinsic calibration in learning-based calibration methods are to ascertain the optical center and focal length. the exterior calibration of objectives, however, is distinct. Provide the translation vector and rotation matrix.

3.1. Calibration of Intrinsic Parameters

Deep focal [19], is a groundbreaking study in machine learning-based calibration. Its objective is to accurately determine the focal length of each image taken under real-world conditions. Deep focal meticulously analyzed a basic pinhole camera model and utilized a deep convolutional neural network to predict the horizontal field of vision. The relationship between the focal length f and horizontal field of vision H_0 can be expressed by the width w of an image.

$$H_0 = 2 \arctan \left(\frac{w}{2f} \right) \quad (1)$$

The calibration characteristics of the camera can over time change because of component wear, temperature variations, or external disturbances such as collisions. To achieve this objective, MisCaliDet [27], suggested a method to determine whether a camera requires intrinsic recalibration. MisCaliDet introduced a novel scalar metric called the average pixel position difference (APPD) to quantify camera miscalibration. Unlike traditional intrinsic parameters like focal length and image center, APPD measures the average difference in pixel positions across the entire image.

3.2 Calibration of Extrinsic

Extrinsic calibration is the process of determining the spatial relationship between a camera and the 3D scene it is capturing, in contrast to intrinsic calibration. PoseNet [28], originally employed deep convolutional neural networks to achieve real-time regression of the 6-DoF camera position. The PoseNet technique produced a pose vector, labeled as p , that includes the position as 3D (x) and orientation (quaternion q) of the camera. Mathematically, the equation can be expressed as $p = [x; q]$. The training dataset is generated by employing a structure from motion technique to automatically compute the labels from a video of the scenario [29].

Expanding on the work of PoseNet [30], further research has improved extrinsic calibration by specifically addressing factors such as intermediate representation, interpretability, data format, and learning aim. For example, Deep FEPE [31] created a sophisticated system that uses key points to detect, extract features, match, and reject outliers. This method aims to improve the accuracy of determining

the geometric pose. This pipeline duplicates the traditional baseline, in which the intermediate differentiable module enables the analysis and enhancement of the final performance.

To resolve the difference between the external objective and image characteristics, recent research has proposed utilizing a representation obtained from the input, such as a depth map, surface geometry, normal flow, and directional probability distribution, among other options. Subsequently, the extrinsic factors are established by considering geometric limitations and obtained representation. As a result, the neural networks are methodically trained to identify the geometric features that are crucial for determining exterior measures. Due of concerns about privacy and limited storage capacity, newer studies have compressed the scene and utilized point-like features to estimate the extrinsic. For instance [32], conducted training on a network to identify sparse yet important 3D locations, referred to as scene landmarks, by representing their appearance as implicit characteristics. The camera pose can be determined by applying a reliable minimum solver, followed by a nonlinear refinement using the Levenberg-Marquardt method. Scene Squeezer utilizes a three-level approach to compress scene information. Firstly, it clusters the database frames based on pairwise co-visibility information. Then, a point selection module prune each cluster by considering estimation performance. Finally, the picked points are further compressed using learned quantization.

3.3. Calibration of both extrinsic and intrinsic parameters simultaneously

3.3.1. Representations Of Geometric

Points of disappearance the convergence of the projections of a collection of parallel lines in the globe results in a vanishing point. Identifying vanishing spots is a fundamental and essential task in 3D vision. Typically, vanishing points indicate the orientation of 3D lines, enabling the observer to infer 3D scene details from a 2D image. DeepVP is the initial study that uses machine learning to detect the points in one image. The process deviated from the usual approach by evaluating the potential horizon lines based on the vanishing spots they encompass.

Chang et al. [33], modified this challenge to be a CNN classification issue by utilizing an output layer that can identify 225 distinct potential vanishing point positions. To create the dataset, the camera is moved horizontally and vertically in increments of 5° , ranging from -35° to 35° , capturing a total of 225 photos of the panoramic scene from a single GPS location. NeurVPS introduced an authorized conic space and a conic convolution operator to exploit the geometric properties of vanishing points. These convolutions can be performed regularly in this space.

The learning model is able to compute the overall geometric information of vanishing points at a local level. To address the issue of requiring a substantial amount of training data in earlier approached [34], a neural network by two different geometric priors: Gaussian sphere and Hough transformation. First, the features are transformed into the Hough domain, where lines are allocated to distinct bins. The Hough bins are mapped onto the Gaussian sphere, where lines are transformed into sizable circles while the vanishing points are situated at the point of intersection of these circles. Geometric priors are advantageous in terms of data utilization as they obviate the necessity of acquiring this information from data. This enables the implementation of a learning framework that is easily understandable and exhibits high performance in areas where there are minor differences in data distributions. Lines that are parallel to the horizon The horizon line is crucial in offering context for various computer vision applications, including picture metrology, computational photography, and the interpretation of 3D scenes. The horizon line is established by projecting the line at infinity onto a plane that is perpendicular to the gravity vector.

Determining the position of the horizon line in the captured image may be easily accomplished by utilizing the camera's Field of View (FoV), pitch, and roll. DeepHorizon introduced the initial learning-based approach to estimate the horizon line from an image, without the need for explicit geometric constraints or other clues. In order to train the network, a novel benchmark dataset called Horizon Lines in the Wild (HLW) was created. This dataset

comprises of real-world photos that have been annotated with labeled horizon lines. SAMobileNet introduced a method for detecting and correcting image tilt using self-attention Mobile-Net [35], specifically designed for smart mobiles. A module of self-attention was created to acquire knowledge of distant relationships and overall context within the input visuals. In order to tackle the challenge of the regression problem, the network was trained to predict numerous angles that fall within a small range around the actual tilt value. Only the values that lie outside of this narrow range were penalized.

3.3.2. Composite Parameters

The process of calibrating the composite parameters involves the simultaneous estimation of both the intrinsic and extrinsic parameters. The study of Hold-Geoffroy et al. [11], achieved superior results compared to earlier independent calibration tasks by simultaneously computing composite parameters with a dataset large-scale [36]. In addition [11], conducted a study on human perception, where participants were tasked with assessing the authenticity of 3D objects that were either composited with or without precise calibration.

This data was subsequently utilized to develop a novel perceptual metric for quantifying calibration

mistakes. Regarding the feature category [37], CTRL-C took into account both geometric cues and semantic characteristics for calibration.

They demonstrated how utilizing geometric cues can help the network understand the fundamental perspective structure of a picture. The process of copying text using the CTRL-C command is depicted in Fig. 3. Recent literature has explored several applications that are examined in conjunction with camera calibration. These applications include single view metrology, shape estimation, 3D human pose, object pose estimation, depth estimation, and image reflection.

CPL utilized a camera model neural network to estimate camera parameters, taking into account the diverse nature and visual subtleties of these characteristics. This was achieved by the use of a unique camera projection loss, which facilitated the reconstruction of the 3D point cloud. The proposed solution aimed to resolve the training imbalance issue by quantifying various faults of camera parameters using a standardized measure.

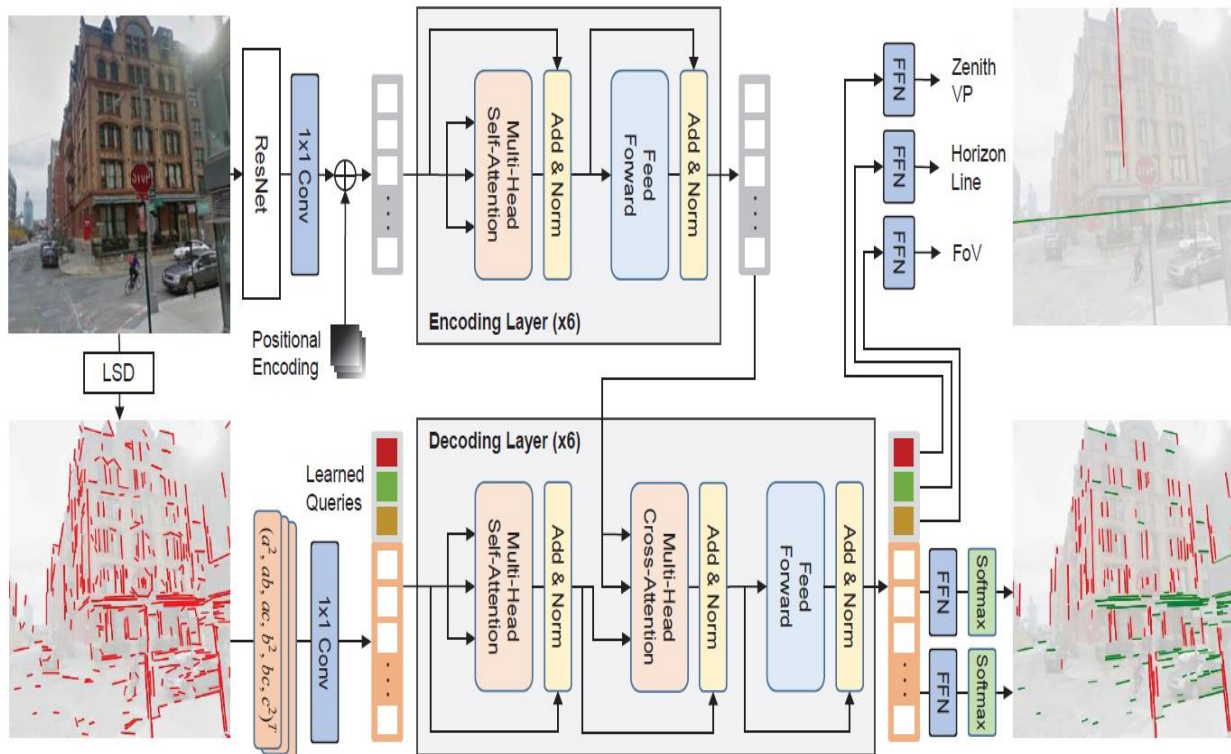


Fig. 3 Overview of CTRL-C [38].

3.4 Discussion

The aforementioned strategies aim to achieve automatic calibration without any manual intervention. Direct intervention and assumption of the situation. In the field of early literature, two distinct studies [21, 22] focused on intrinsic calibration and extrinsic calibration. Building upon extensive datasets and robust networks, later studies [11, 40, 41], have explored a whole camera calibration process, deducing several parameters and geometric representations. To alleviate the challenge of acquiring knowledge about the camera settings, certain studies [42-44], have suggested the idea of acquiring an intermediary representation through learning. Recent literature has explored the simultaneous investigation of camera calibration with various applications [45-49]. This implies that addressing the vision problems that occur later in the processing pipeline, particularly those involving three-dimensional tasks, may necessitate familiarity with the model used to create the images. In addition, some geometric assumptions [50] can reduce the need for large amounts of data in deep learning,

indicating the possibility of connecting the calibration objective with semantic information.

It is intriguing to discover that the implementation of additional extrinsic calibration methods [51,48], has led to a reevaluation and reinstatement of the conventional feature point-based solutions. The camera motion is described by a set of extrinsic that have a restricted number of degrees of freedom. As a result, some features may effectively express the spatial correspondence. In addition, the network specifically developed for point learning greatly enhances the effective models of calibration, such as Point-CNN and Point-Net This pipeline also facilitates the clear interpretation of camera calibration based on learning, which enhances the comprehension of how the network adjusts and amplifies the effects of intermediary modules.

- (1) Investigate additional visual and geometric assumptions. Given the limited availability of real-world data in the field of learning-based calibration, it is promising to explore additional prior knowledge that can reduce the reliance on data-driven learning. For instance, the prior image creation model enables us to establish the

connection between the settings of the 3D camera and the image layout of the 2D camera.

- (2) Separate the various stages for an end-to-end model of calibration. Typically, calibration approaches that rely on learning involve two main stages: feature extraction and objective estimation. Nevertheless, the process by which networks acquire knowledge about calibration-related properties remains unclear. Thus, separating the learning process through many conventional calibration phases helps direct the process of feature extraction. Expanding the concept of extrinsic calibration [51-53], to encompass a wider range of calibration difficulties would have significant significance.
- (3) Change the measurement value space from geometric difference to parameter error. The training procedure will encounter an imbalance loss optimization challenge when attempting to jointly calibrate several camera settings. distinct camera parameters correspond to distinct sample distributions, which is the fundamental reason. The basic normalization approach is unable to merge their error spaces. Thus, we may establish a clear and precise measuring framework based on the geometric characteristics of various camera parameters.

4. Distortion Camera Model

Camera calibration using machine learning techniques is gaining significant interest for its ability to accurately calibrate radial distortion and roll

shutter distortion. These distortions are particularly important in wide-angle applications. Optical lens with complementary metal-oxide-semiconductor (CMOS) sensor. In this section, our primary focus is on examining the rectification and calibration of both aberrations.

4.1 Radial Camera Distortion

It refers to distortion that occurs in an image when the distance from the center of the image increases. The existing body of research on learning based on radial camera distortion calibration can be divided into two primary categories: reconstruction-based solutions and regression-based solutions.

4.1.1 Solution Based on Regression Analysis

Deep Calib in [23] and [54] are groundbreaking studies in the field of wide-angle camera calibration using machine learning techniques. The calibration was approached as either a supervised regression [54], or classification [23], problem, and thereafter, the networks were utilized.

The convolutional and fully connected layers were employed to acquire knowledge about the distortion characteristics of inputs and make predictions about the parameters of the camera. Deep Calib [54], specifically investigated three learning approaches for calibrating wide-angle cameras, as shown in Fig. 4. Their trials demonstrated that the Single Net, with its simplest architecture, achieved the highest level of performance in terms of both accuracy and efficiency.

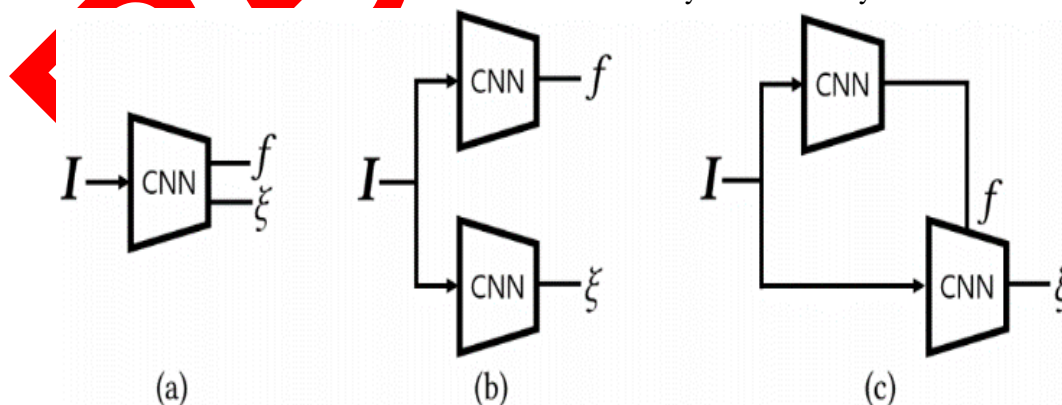


Fig. 4 The regression-based wide-angle camera calibration: (a) Single Net, (b) Dual Net, and (c) Seq Net [59].

Where I represent the distorted image, f , and ξ represent the focal length and distortion parameters

In order to improve the perception of distortion in networks, previous studies have explored the incorporation of a wider range of features, including semantic features and geometric features [55- 58]. In

addition, certain studies enhanced the ability to generalize by implementing learning techniques such as unsupervised learning [59], self-supervised learning [61], and reinforcement learning [60]. By employing randomly selected coefficients during each mini-batch of the training procedure, RDC-Net [53], was capable of dynamically generating distortion images in real-time. It improved the accuracy of the rectification process and mitigated the risk of the learning model overfitting. Instead of making contributions to the advancements in deep learning algorithms, other research focused on exploring the visual aspects before achieving interpretable calibration. For instance, PSE-GAN [62], and [63], developed a position-aware weight layer that takes into account the radial distortion characteristics of an image, specifically the stronger distortion observed in the texture far from the image center. This layer, which can be either fixed [63], or learnable [62], allows the network to explicitly perceive and account for the distortion. In study of Lopez et al. [64], introduced a new method of parameterizing radial distortion that is more suitable for networks than directly learning the distortion parameters.

In addition, Ordinal Distortion [65], introduced a representation called ordinal distortion, which is conducive to learning. In contrast to the implicit and varied parameters of camera, this representation can enhance the neural network's ability to perceive distortion by establishing a clear connection with the visual elements.

4.1.2 Solution based on reconstruction

The reconstruction-based method is inspired by visual image-to-image dense perception and translation. It represents a departure from the

traditional regression-based approach [32]. By explicitly simulating the pixel-by-pixel mapping between the distorted and rectified images, DR-GAN is a revolutionary technique to radial distortion calibration. One-stage correction and free camera parameter training were accomplished during the training phase. The liberation from the assumption of camera models has made it possible to create a reconstruction-based method that is able to calibrate multiple kinds of cameras in one learning network. For example, DDM [33], introduced the distortion distribution map, which allowed many camera models to be combined into a single domain. For every pixel in a warped image, this map shows the exact amount of distortion. The network then learned to reconstruct the corrected image using the geometric previous map.

Several subsequent studies [35, 60, 66, 67], concentrated on finding the displacement field between the distorted picture and the rectified image in order to improve the interpretability of the mapping function. It is possible to eliminate the artefacts created during the pixel-by-pixel reconstruction of the image using this method. The geometry prior from Shi et al. and PSE-GAN was included into FE-GAN's reconstruction-based methodology. As seen in Fig. 5, they also presented a self-supervised method for learning the distortion flow for wide-angle camera calibration. A U-Net-like architecture is used by several reconstruction-based systems to learn about pixel-level mapping. However, the distortion feature might be transmitted from the encoder to the decoder via the skip-connection procedure, which would lead to a hazy look and insufficient correction in the reconstructed results.

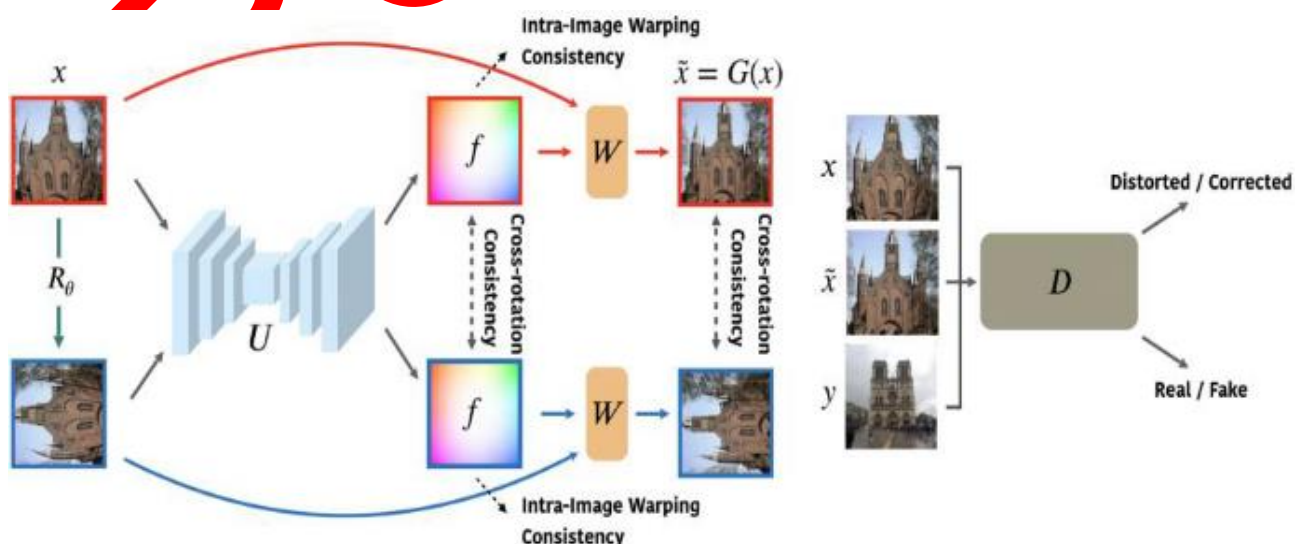


Fig. 5 Architecture of FE-GAN [68].

In order to resolve this problem, Li et al. [69], decided to eliminate the skip connection in the rectification network. In order to simultaneously maintain feature fusion and control geometric differences, PCN [70]. implemented a correction layer within the skip-connection and utilized appearance flows to modify the convolved features in various encoder layers. PolarRecNet [71], addressed the issue of the prior sampling approach of the convolution kernel by taking into account the radial symmetry of distortion. It achieved this by transforming the deformed image from the Cartesian coordinate's domain to the polar coordinate's domain.

4.2 Analysis of Roll Shutter Distortion

The current deep learning calibration for roll shutter (RS) distortion can be divided into two model: single-frame-based [72, 73] and multi-frame-based [74]. The numbers 150, 155, 156, and 163 are listed. The single-frame-based technique focuses on analyzing a single image roll shutter as input and utilizes neural networks to immediately learn and correct the distortion. The optimal outcome can be considered as the global shutter (GS) image. The problem is poorly defined and necessitates the establishment of additional prior assumptions.

In contrast, the multi-frame-based method takes into account the consecutive frames of video captured by a roll shutter camera. This allows for the investigation of the significant temporal connection, leading to a more appropriate correction.

4.2.1 Solution based on single frames

URS-CNN [72] is the initial study that focuses on the calibration of rolling shutter cameras through learning. This study employed a neural network with extended kernel properties to investigate the relationship between picture structure and row-wise camera movements. In order to explicitly examine the RS effect caused by the row-wise exposure, we utilized row-kernel and column-kernel convolutions to extract characteristics along the horizontal and vertical axes. RSC-Net [75] enhanced the URSCNN [72], by increasing the degrees of freedom (DoF) from 2 to 6, and introduced a correction model that is aware of the structure and motion of the remote sensing (RS) data. This model estimates the velocity of the camera scanline and the depth information.

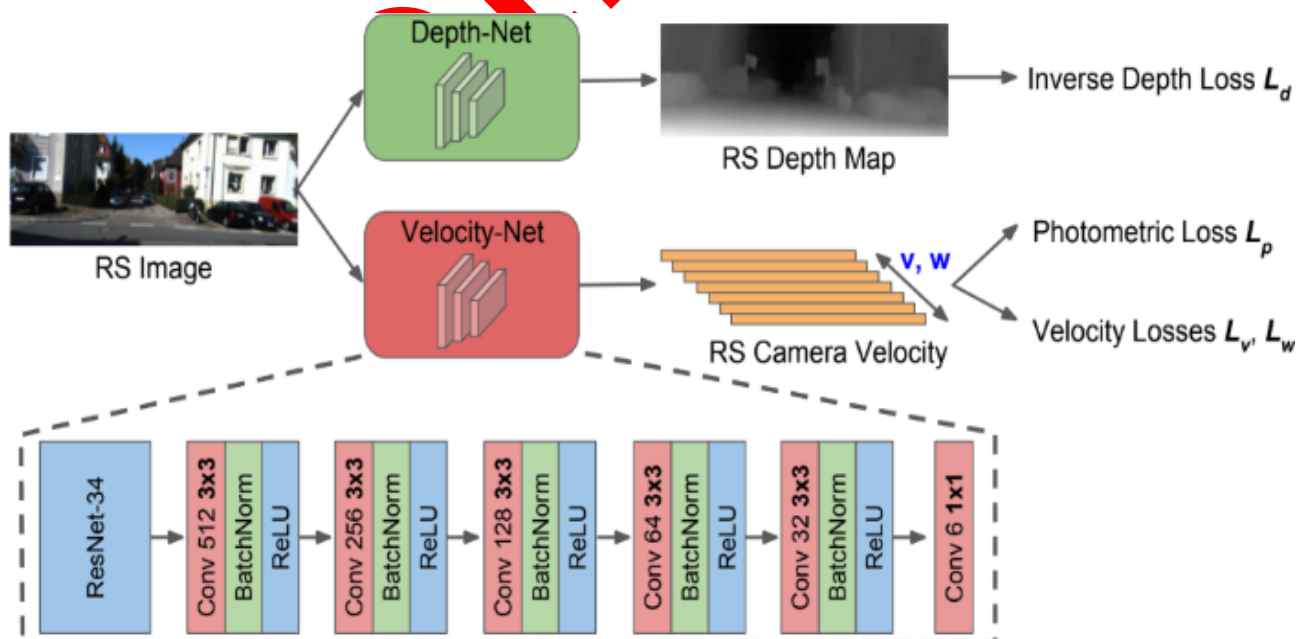


Fig. 6 Architecture of RSC-Net [75].

In contrast to RSC-Net [75], URS-CNN [72], additionally considered the hidden motion between the scanlines and the structure of the picture, as depicted in Fig. 6. To establish a relationship between RS and GS, EvUnroll [73] utilized neuromorphic events to rectify the RS effect. Event cameras can mitigate certain limitations of traditional frame-based cameras for capturing fast-moving dynamic scenes, thanks to their microsecond-level sensitivity and high temporal resolution.

4.2.2. Solution of multi-frame-based

Most multi-frame-based systems adhere to the reconstruction paradigm, which focuses on warping the RS domain to the GS domain properly and capturing the dense displacement field between global GS and RS pictures. Using a differentiable module of forward warping, Deep Unroll Net [74] presented an end-to-end network that manages two consecutive rolling shutter photographs.

In order to precisely determine the dense displacement field between a rolling shutter image and its equivalent global shutter image, this method makes use of a motion estimation network. Another contribution comes from DeepUnrollNet [74], which generates two new datasets: the Carla-RS dataset and the Fastec-RS dataset. Furthermore, JCD [76] investigated the combined use of deblurring (RSCD) and rolling shutter correction (RSCD) techniques, which are primarily employed when rolling shutter cameras are utilized for medium-to long exposures. In order to compensate for displacement and maintain the nonwrapped deblurring stream for detail restoration, the technique employed bi-directional warping streams. Furthermore, the authors offered a useful dataset that made use of the painstakingly constructed beam-splitter acquisition technique known as BS-RSCD. This dataset includes item movement in dynamic situations as well as the movement of the observer.

SUNet [69] expanded Deep Unroll Net [74] by using the intermediate time between two frames starting from the midway time of the second frame. The symmetric un distortion fields were estimated and the possible GS frames were rebuilt using PWC-Net [80] and SUNet [77], respectively, through a

time-centered GS image decoder network. The context-aware un distortion flow estimator and symmetric consistency enforcement were developed to successfully minimize the misalignment between the distorted contexts of two consecutive RS images. To increase the frame rate [78], created a GS video by utilizing the scanline-dependent characteristics of the RS camera to combine two consecutive RS images. More precisely, they started by looking at the intrinsic connection between optical flow and bidirectional RS un distortion flow. The RS distortion flow map, as opposed to the isotropic smooth optical flow map, demonstrated a greater reliance on scanlines. The researchers then designed bidirectional Un-distortion flows to express the displacement at the pixel level that is cognizant of the remote sensing (RS) information.

They also established a computational method to convert between distinct RS un distortion flows for different scanlines. In order to address the issues of inaccurate displacement field estimate and error-prone warping in prior methods, AW-RSC [79] suggested a solution that involves predicting multiple fields and adaptively warping the learnt RS features into global shutter equivalents. By employing a method that progresses from a rough to a detailed approach, the distorted characteristics were merged and produced into accurate global shutter frames, as depicted in Fig. 7. In contrast to prior studies [74,58,62,49], the warping operation in AW-RSC [80] is both trainable and efficient, thanks to the inclusion of adaptive multi-head attention and a convolutional block. Furthermore, AW-RSC [81] provided a dataset specifically designed for correcting the rolling shutter effect in real-world scenarios. BS-RSC refers to a system where RS movies and their associated GS ground truth are collected at the same time using a based acquisition beam-splitter technique.

4.3 Analysis and conversation

4.3.1 Methodology Overview

Deep learning algorithms are designed to operate on images captured by wide-angle cameras and account for the roll of the camera. The process of calibrating shutters involves using a similar approach pipeline.

In line with this research tendency, the majority of early literature commences with the solution based on regression [82, 83, 72]. The following works revolutionized the conventional calibration by adopting a reconstruction approach [84, 85], which directly estimates the displacement field to correct the uncalibrated input. To enhance the precision of

calibration, researchers have devised a more displacement field and efficient warping technique [49, 70, 86]. In order to accommodate the various types of distortions, certain studies have created convolutional kernels with varying forms [72] or have altered the coordinates of the convolution [71].

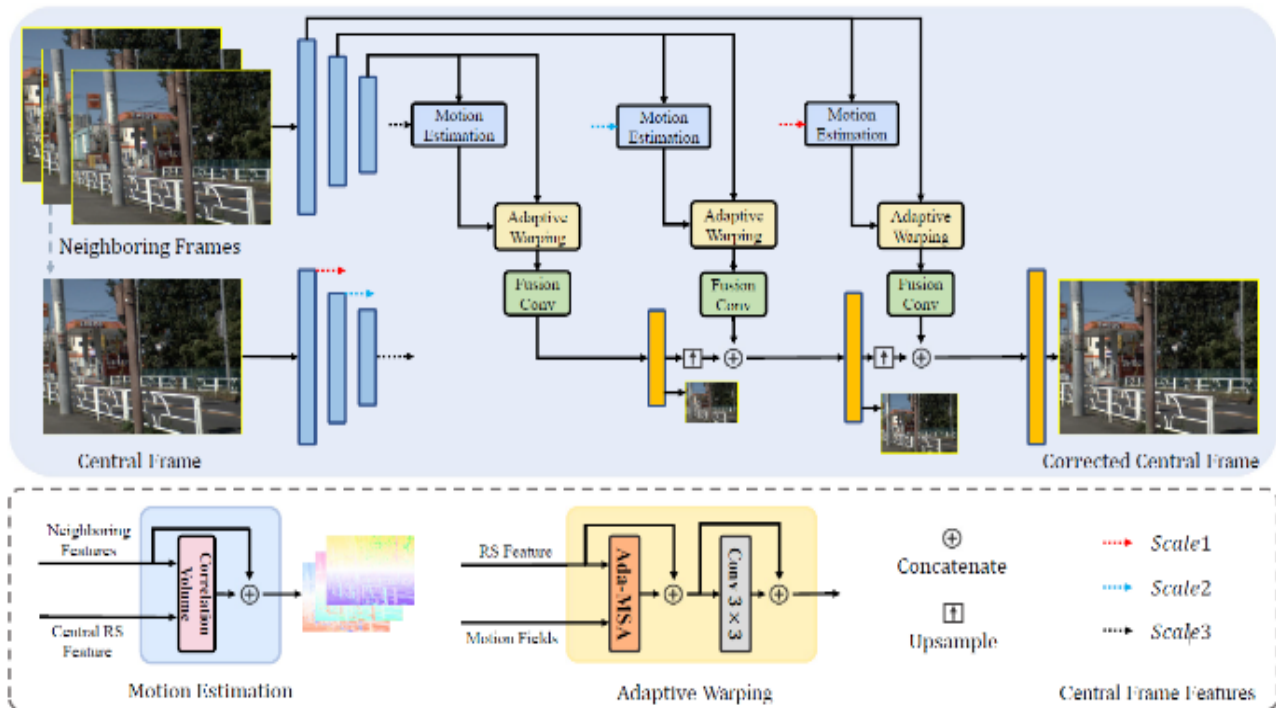


Fig. 7 Architecture of AW-RSC [87].

Prior research focused on developing more robust networks and including a wider range of features to enhance calibration performance. There is a growing trend in using geometric priors to address distortion, as seen in methods such as [63,62, 68]. The priors can be incorporated into the convolutional layers or utilized to supervise the training of the network, hence enhancing the speed at which the learning model converges.

4.3.2 Subsequent Endeavors

(1) The advancement of roll shutter camera calibration and wide-angle camera calibration can mutually enhance one another. An example of a well-researched approach in roll shutter calibration is the multi-frame-based solution, which can effectively facilitate wide-angle calibration.

Objects positioned at various sequences but with the same characteristics can offer valuable previous information regarding radial distortion. In addition, the comprehensive investigations of the warping layer and displacement field [76, 87] have the capacity to inspire the advancement of camera calibration wide-angle and other related areas. Additionally, exploring geometric priors in wide-angle calibration could enhance the comprehensibility of the network in shutter calibration.

(2) The majority of approaches create their training dataset by combining random samples from all camera parameters. However, the distribution of camera parameters for images acquired by real lenses is likely to be located on a possible manifold [64]. Training on a calibration dataset that has duplicate labels hinders the efficiency of the learning process. Hence, it would be

worthwhile to investigate a pragmatic sampling approach for the generated dataset as a potential future endeavor.

- (3) To address the difficulty of single-frame calibration, incorporating additional high-precision sensors, such as event cameras [73], can enhance the
- (4) existing calibration performance. The advancement of visual sensors has led to the value of joint calibration employing several sensors.

5. The Cross View Model

Current deep calibration techniques are capable of accurately determining the precise camera parameters using only a single camera. Indeed, in multi-camera settings, parameter representations might become more intricate. In the Multiview model, the basic matrix and essential matrix are used to characterize the epipolar geometry. These matrices are closely connected to the intrinsic and extrinsic parameters. The homography represents the pixel-level correspondences between distinct viewpoints. Furthermore, depth is intricately connected to both intrinsics and extrinsic. Out of all the complicated ways to describe parameters, homography is the most commonly used in practical applications and has been extensively studied in relation to learning-based methodologies. In order to achieve this objective, our main focus is on examining several approaches for estimating deep homography in this model. These approaches can be categorized into three main groups: cascaded, direct, and iterative solutions.

5.1 Direct Model Solution

We analyze the direct homography solutions by examining several parameterizations, such as the traditional 4-point parameterization and alternative parameterizations.

5.1.1. 4-pt model Parameterization

Deep homography estimation was first introduced in DHN [88], where a VGG-style network is utilized to predict the 4-point parameterization H_{4pt} . To train and evaluate the network, a synthetic dataset called Warped MS-COCO was created to

provide ground truth for the 4 – point parameterization \hat{H}_{4pt} . The pipeline is shown in Fig. 8(a), and the objective function is defined as LH.

$$L_H = \frac{1}{2} \|H_{4pt} - \hat{H}_{4pt}\| \quad (2)$$

Then 4-point parameterization can solve using a 3×3 matrix. The homography matrix is computed using the normalized Direct Linear Transform (DLT) algorithm [88]. Nevertheless, DHN constrained to synthetic dataset in which the ground truth may be constructed without expense or necessitates expensive labelling of real-world datasets. Later on, a solution called UDHN [89] is presented to tackle this problem without the need for supervision. demonstrates that it employed the identical network structure as DHN and established an unsupervised-loss function by minimizing the average of photometric error, which was inspired by conventional techniques [90]:

$$L_{HPW} = \|P(I_A(x)) - p(I_B(W(x; p)))\| \quad (3)$$

$W(x; p)$ and $P(I)$ represent the actions of warping.

By utilizing homography parameters (P) and extract image patch, the desired outcome can be achieved. I_A and I_B are the initial photos that have areas that overlap with each other. UDHN takes a pair of picture patches as input, however it distorts the images while computing the loss. By following this approach, it prevents the negative consequences of distorted pixels and enhances the level of pixel oversight. In order to enhance both speed and accuracy of a small-scale model, Chen et al. introduced Shuffle-Homo-Net [91], which combines Shuffle-Net compressed [92] and location aware [77] into a new model (lightweight) as displayed in Fig. 8.

In order to effectively deal with significant changes in position, a form of weight-sharing that operates at several scales is utilized. This involves extracting feature representations at different scales and then intelligently combining predictions from these different scales. Nevertheless, the homography is unable to achieve a flawless alignment of pictures due to

parallax resulting from non-planar structures and non-overlapping camera centers. In order to address the parallax issue, CA-UDHN [93] employs trainable attention masks to disregard the parallax regions, hence improving the

alignment of the background plane. In addition, the 4-point homography can be expanded to mesh-flow [94] in order to achieve precise alignment of non-planar objects as shown in Fig. 9.

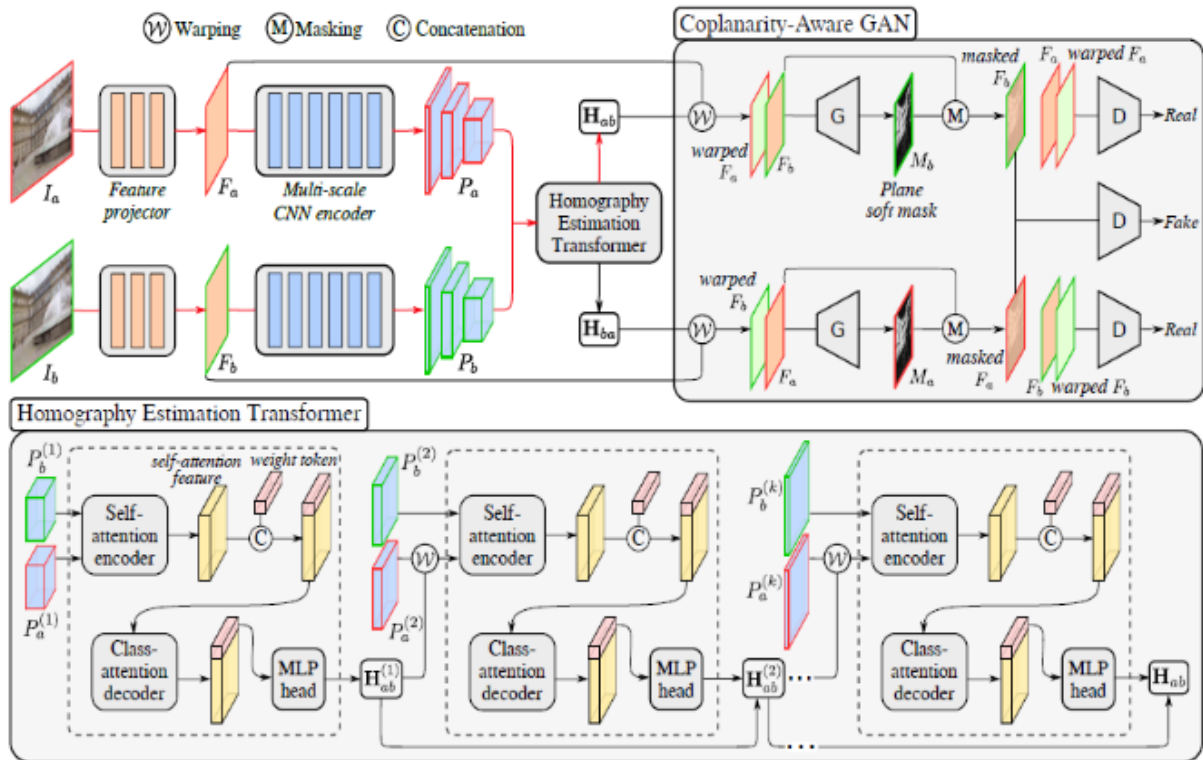


Fig. 8 The Architecture of HomoGAN [94].

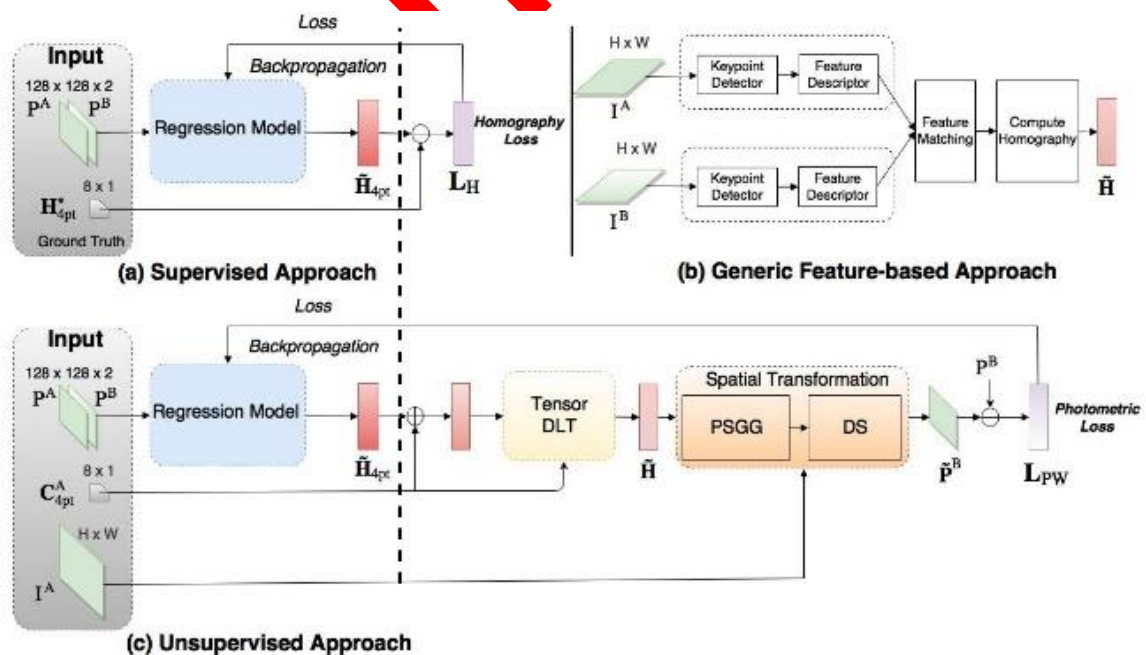


Fig. 9 Architectures of DHN [25] and UDHN [88]

5.1.2 Alternative Parameterizations

Aside from the 4-point parameterization, the homography can also be expressed using alternative formulations. Wang et al. introduced SSR-Net [95] to enhance the utilization of homography invertibility. The invertibility requirement was imposed using a normal matrix represent in a cyclic fashion. Zeng et al. [96] contended that using a fully-connected layer to regress the 4-point parameterization can disrupt the spatial arrangement of the corners and make it vulnerable to disturbances, as it involves four points. Figure 9 displays the architectural design of Homogamy. The source of the figure is cited as [94]. Points are the essential prerequisite for solving the homography. To tackle these problems, they devised a parameterization technique known as a perspective field (PF), which represents a one to-one mapping between pixels. Additionally, they developed a PFNet to implement this technique.

This process expands the positions of the four vertices to include as many closely spaced pixel points as feasible. The homography can be determined by applying RANSAC [97] with outlier filtering, which allows for reliable estimation by utilizing dense correspondences. However, the presence of several correspondences results in a substantial rise in the computational intricacy of RANSAC. In addition, Ye et al. [98] introduced an 8-DOF flow representation that does not require any post-processing. This representation has a size of $H \times W \times 2$ and is bound by the homography in an 8-dimensional subspace.

To describe arbitrary homography flows within this subspace, a total of 8 flow bases have been created. The suggested method, BasesHomo, aims to predict coefficients for these flow bases. To acquire the desired bases, BasesHomo initially generates 8 homography flows by altering each individual element of an identity homography matrix, except for the last entry. Next, the flows are normalized based on their highest flow magnitude and then subjected to a QR decomposition to ensure that all the bases are both normalized and orthogonal.

5.2 Solution of cascaded

Direct solutions investigate several ways of parameterizing homography using straightforward network structures, whereas the cascaded approach One's attention is directed towards intricate network architecture designs. In the Hierarchical Net model proposed [99], it is argued that distorted images can be considered as input for another neural network. Consequently, they arranged networks in a sequential manner to decrease the error limits of the estimation. SRHEN [100] incorporated the volume into the cascaded network, utilizing cosine distance to measure the feature correlation and representing it as a volume. While the cost volume and stacked networks do enhance performance, they are unable to effectively handle dynamic scenarios. MHN [101] created a neural network that operates at several scales and suggested training it to learn both homography estimation and dynamic content detection at the same time.

In addition, Local Trans [37] addressed the issue of cross-resolution by formulating it as a multimodal problem. They suggested a local transformer network that is integrated into a multiscale structure to effectively learn the connections between the multimodal inputs. The inputs consist of photos with varying resolutions, and Local Trans demonstrated exceptional performance in circumstances where there was a resolution difference of up to 10 times. All the a forementioned strategies utilize image pyramids to gradually improve the capability to handle significant displacements. Nevertheless, each image pair at every level necessitates a distinct extraction feature network, leading to duplication of feature maps. To address this issue, certain researchers [102, 103, 94] substituted images with the feature pyramids. More precisely, they manipulated the feature maps directly instead of photos to prevent the use of excessive feature extraction networks. To tackle the issue of estimating homography with minimal overlap in real-world photos, Nie et al. [103] made modifications to the unsupervised constraint (Eq. 4) to make it suitable for low-overlap situations.

$$L_{pw} = \|I_A(x) \cdot I(W(x; p)) - I_B(W(X; p))\| \quad (4)$$

Where 1 is a matrix filled with all ones, having same size as (IA or IB). To address the issue of low-overlap, the solution involved using the original photos as input for the network and removing the related elements.

Mapping the pixels of image A to the pixels of image B that have been distorted. DAMGHomo [104] offered a solution for the challenge of estimating non-planar homography. They proposed a method called reverse multi-grid deformation with contextual correlation to align parallax pictures. The proposed contextual correlation method achieved higher accuracy with reduced computing complexity compared to standard cost volume. An alternative approach to tackle the non-planar issue is to concentrate on the prevailing plane. HomoGAN, introduced in [94], presents an unsupervised GAN that applies a coplanarity constraint to the predicted homography, as depicted in Figure 9. To execute this method, a generator is employed to forecast masks of aligned regions, while a discriminator is utilized to ascertain whether two masked feature maps were generated by a single homography.

5.3 Solution using iteration

Iterative solutions produce improved accuracy compared to cascaded methods by repeatedly optimizing the final estimation. The (LK) algorithm Lucas-Kanade is commonly employed in picture registration to repeatedly estimate parameter warps, like affine transformations and optical flow. The objective is to iteratively update the warp parameters by minimizing the sum of squared error between a template picture T and an input image I.

$$E(\Delta p) = \|T(x) - I(W(x; p) + \Delta p)\| \quad (5)$$

However, when optimizing Equation 5 using a first-order Taylor expansion, it is necessary to recompute $\delta I(W(x; p)) = \delta p$ at every iteration because the value of $I(W(x; p))$ changes with p. To prevent this from happening. The problem can be addressed by using the inverse compositional (IC) LK method, which is an equivalent alternative to the LK algorithm. This technique allows for the reformulation of the optimization aim in the following manner:

$$\hat{E}(\Delta p) = \|T(W(x; \Delta p)) - I(W(x; p))\| \quad (6)$$

By applying a first-order Taylor expansion to Equation 6 and linearizing it, we obtain the expression $\Delta T(W(x; 0)) = \Delta p$ instead of $\Delta I(W(x; p)) = \Delta p$. This substitution ensures that the value does not change with each iteration. To use the benefits of deep learning in conjunction with IC-LK iterator, CLKN [105] performed LK iterative optimization on semantic feature maps that were extracted by CNNs. The process is as follows:

$$E^f(\Delta p) = \|F_T(W(x; \Delta p)) - F_T(W(x; p))\| \quad (7)$$

FT and FI represent the feature maps of the template and Images provided as feedback. Subsequently, they compelled the network to operate a Perform a singular iteration using a hinge loss function while the network is running repeatedly iterate until the specified stopping condition is satisfied. the phase of testing. In addition, CLKN arranged three identical LK in a pile. Networks can be utilized to enhance performance by addressing the issue. The output of the previous LK network serves as the first warp settings. of the upcoming LK network. According to the equation. The IC-LK algorithm is denoted as 7. The system mainly depends on feature maps, which have a tendency to be ineffective in multimodal scenarios. Visual representations. DLKFM [106] built a single-channel instead. Generate a feature map by utilizing the eigenvalues of the local data. The output tensor's covariance matrix. To acquire knowledge of DLKFM, one must I developed two unique constraint terms to synchronize multimodal elements. Feature maps play a role in the process of convergence.

Nevertheless, algorithms based on LK may encounter difficulties in the presence of an unreliable Jacobian. The matrix has a rank deficiency, as indicated by the value 194. In addition, the IC-LK iterator Is not capable of being trained, indicating that this limitation is purely theoretical inevitable. To resolve this matter, a comprehensive the trainable iterative homography network (IHN) [57] was developed. suggested. Building upon the concept of RAFT, IHN modifies the cost. Enhance the approximated homography by adjusting the volume accordingly.

The estimator is repeatedly executed during each iteration. In addition, IHN can

Address dynamic scenarios by generating an inlier mask inside the given context. Autonomous estimator that operates without the need for additional oversight.

5.4 Discussion of Model

5.4.1 The Summary

The aforementioned works focus on investigating several methods of parameterizing homography, including the perspective field [82], 4-point parameterization [88], and motion bases representation [98]. It enhances performance and convergence. The other works typically aim to create diverse network architectures. Specifically, the proposed techniques involve using cascaded and iterative methods to gradually improve performance. These methods can also be coupled to achieve even greater precision. To enhance the applicability of the approaches, several difficult issues are initially tackled. These include multiple modalities, cross resolutions [107,108,57], and non-planar sceneries [102], dynamic objects [92,94,101], among others.

5.4.2 Future Effort and Challenge

The existing challenges can be summarized as following:

- (1) Numerous homography estimate algorithms are specifically tailored for static resolutions, although real-world applications frequently necessitate significantly more adaptable resolutions. Applying pre-trained models to images with varying resolutions might result in a significant decrease in performance. This is because the images need to be resized to meet the required resolution, which can negatively impact the model's performance.
- (2) In contrast to flow optical estimation, which requires tiny movements between images, estimation homography frequently involves images with relatively low overlap rates. In such instances, the performance of conventional approaches may be subpar because of their restricted receptive fields.
- (3) Current approaches tackle the issue of parallax or moving objects by training models to identify and discard data points that do not fit the expected

pattern in the extractor feature [92], estimator [109], or cost volume [110]. Nevertheless, it remains ambiguous as to whether stage is more suitable for the rejection of outliers.

Given the problems we have discussed, we may identify some prospective study topics for future endeavors:

- (1) To address the initial obstacle, we can develop diverse tactics to improve the ability to handle different levels of resolution. This can be achieved by techniques including augmenting data relevant to resolution and continuously learning from several datasets that have varying resolutions. In addition, we have the option to create a parameterization form that does not require a resolution. The perspective field [96] is an exemplary example that displays the homography by using dense correspondences by same resolution as the import images. However, the use of RANSAC as the post-processing strategy adds additional computational cost, particularly when dealing with many correspondences. Hence, it is imperative to investigate a parameterization form that is both resolution-free and efficient.
- (2) To improve the performance when there is a low-rate overlap, the key idea is to enlarge the networks receptive fields. cross-attention module of the transformer effectively utilizes the long-range correlation to remove any inherent bias towards short-range connections. Alternatively, we can utilize advantageous forms of cost volume to incorporate feature correlation [102].
- (3) Since there is no interaction between various picture features in the feature extractor, it is logical to conclude that outlier rejection should take place after feature extraction.

Identifying outliers within a single image is not feasible since relying solely on depth as an outlier cue is insufficient. Images taken by cameras that are only rotated do not have any parallax outliers. Furthermore, it becomes logical to acquire the skill of identifying and excluding outliers by merging both correlation (global and local), akin to the understanding of RANSAC.

6. Cross Sensor Model

Camera Multi sensor calibration is a process that determines the inherent and external characteristics of several sensors, such as cameras, LiDAR's, and IMUs. This guarantees the synchronization of data from several sensors. and recorded in a shared coordinate system, enabling them to be combined for a more precise depiction of the surroundings. Precise calibration of several sensors is essential for applications such as autonomous driving and robotics, where dependable sensor fusion is required for safe and efficient functioning.

In this section, we primarily examine the existing research on learning-based camera-LiDAR calibration. This involves predicting the 6-DoF body rigid transformation between the 3D-LiDAR and the camera, without the need for any specific features or landmarks during the implementation. Like the process of calibration in other camera systems, this study topic can also be categorized into two categories of solutions: solutions flow-based and solutions regression-based. However, Following the matching notion in camera calibration-LiDAR, we divide the extant learning-based research into three categories: object/key point-level solutions, semantics-level solutions, and pixel-level solutions.

6.1 Solution at the Pixel-Level

Initial deep learning method for camera -LiDAR calibration, known as Reg-Net [111], employed Convolutional Neural Networks (CNNs) to integrate feature extraction, feature matching, and global regression to deduce the 6-DoF extrinsic parameters. The system independently analyzed the RGB and LiDAR depth map, and then split the data into two parallel network streams. Next, a dedicated correlation layer was suggested to convolve the combined RGB and LiDAR characteristics into a unified representation. Following the process of feature matching, the integration of global information and regression parameter was accomplished using two fully layers connected, employing a Euclidean loss function. Inspired by this research, subsequent studies have advanced the field of camera-LiDAR calibration by focusing on various aspects such as geometric constraint [112], loss design [113], extraction feature, fusion feature, matching feature [113,115], and calibration representation [116], [62].

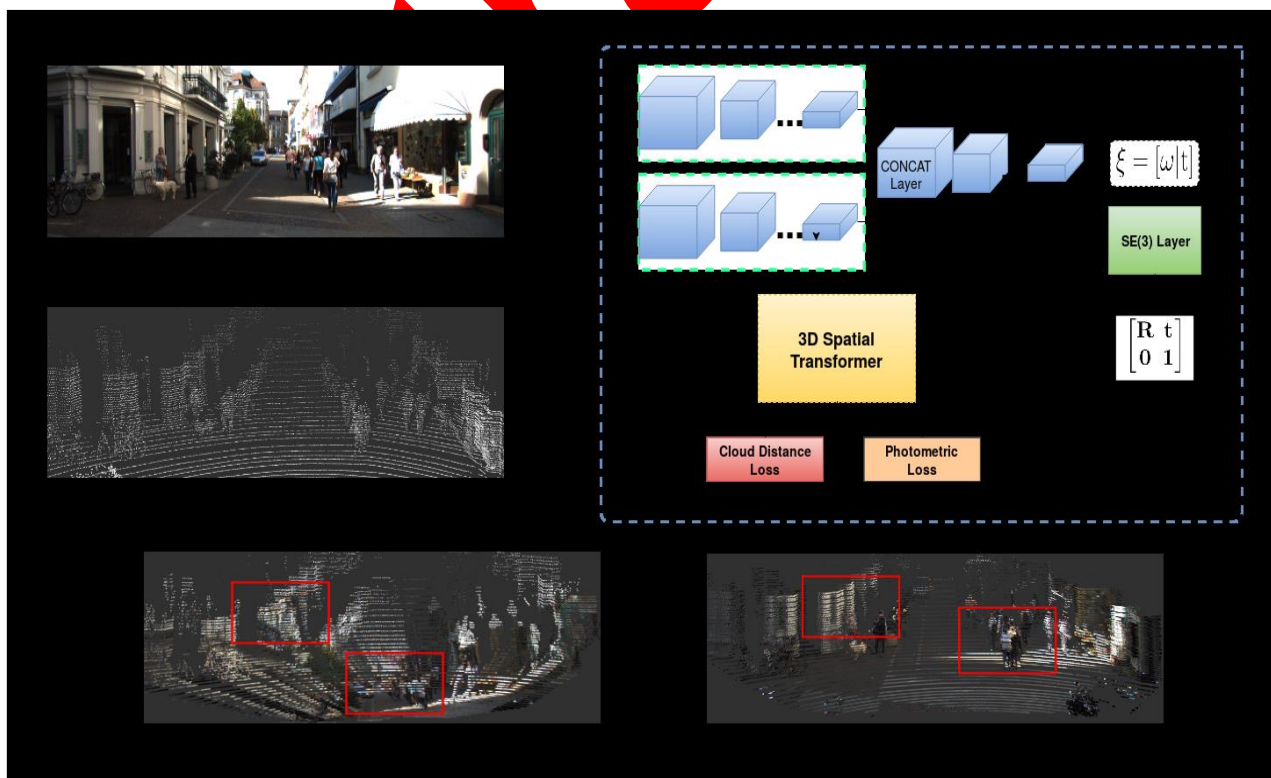


Fig. 10 Network architecture of Calib-Net [112].

Fig. 10 demonstrates the development of Calib-Net [112], a network that predicts calibration parameters to enhance the geometric and photometric consistency of images and point clouds. This is accomplished by employing 3D Spatial Transformers [117] to address the fundamental physical issue. The CalibR CNN [118] method employed a synthetic view and an epipolar geometry constraint to enhance the calibration model. This enabled them to quantify the discrepancies in both the photometric and geometric characteristics between successive frames. In addition, they conducted research on the utilization of LSTM network for acquiring temporal information in camera-LiDAR calibration, which was an innovative methodology. The camera calibration - LiDAR output space is represented in three dimensions using the Special Euclidean Groups (SE(3)), which is distinct from the standard Euclidean space. To tackle this issue, RGGNet [113] integrated Riemannian geometry constraints into the loss function. More precisely, the calibration network was optimized using a SE (3) geodesic distance that employed left-invariant Riemannian metrics.

LCCNet [57] employed the cost volume layer to acquire information regarding the correlation between the image and depth that has been modified by the point-cloud. Fusion-Net [119] employed Point-Net [120] to directly extract characteristics from a 3D point-cloud, as the depth map does not consider the three-dimensional geometric arrangement of the point-cloud. Subsequently, a methodology called feature fusion with Ball-Query and an attention strategy was devised to effectively merge the attributes of photos and point clouds. CFNet originally introduced the calibration pipeline for the aim of calibrating cameras and LiDAR sensors. This flow quantifies the difference between the placements of the initially projected 2D points and the true ground truth. Unlike explicitly anticipating extrinsic factors, the network's comprehension of the fundamental geometric restriction was improved by acquiring knowledge about the calibration flow. To achieve precise 2D and 3D correspondences, CFNet [90]

Corrected the initially projected locations by using the expected calibration flow. Later, the

Efficient-Perspective-n-Point (EPnP) technique was employed to calculate the final extrinsic parameters using the RANSAC method. DXQNet [86] introduced a probabilistic model to the camera calibration flow - LiDAR to address the non-differentiability issue in RANSAC. This model assesses the level of uncertainty to determine the correctness of the camera data - LiDAR association. Later, a posture estimate module with the ability to distinguish was created to solve the problem of determining extrinsic parameters. This module efficiently transmits the external mistake to the flow-prediction network.

6.2 Semantics Solution level

Deep neural networks are capable of effectively learning and representing semantic characteristics. Optimal calibration ensures to precisely synchronize the identical occurrence in various sensors. To achieve this objective, several studies [121, 120, 122] have investigated the use of semantic information to assist in the calibration of camera-LiDAR systems. The SOIC [122] method calibrates and converts the initialization issues to PnP problem of semantic centroids by utilizing semantic knowledge.

A matching constraint cost function was introduced based on the semantic components, as the 3D semantic centroids of the point cloud and the 2D semantic centroids of the picture cannot align perfectly. SSI-Calib [120] redesigned the calibration process by transforming it into an optimization issue. They introduced a new quality metric that relies on semantic properties. Subsequently, a non-monotonic sub-gradient ascent technique was introduced to compute the parameters of calibration. Other studies employed pre-existing segmentation networks for both point cloud and image data. They then enhanced the accuracy of the calibration parameters by minimizing the loss of semantic alignment in both single-direction and bi-directional approaches [51,121].

6.3 Solution at the Object/Key point Level

ATOP developed a Cross-Modal Matching Network called Attention-based Object-Level Matching Network. This network was meant to investigate the overlapping field of view (FoV) between camera and

LiDAR. Its purpose was to assist in producing correspondences between 2D and 3D objects at the object level. The YOLOv4 and Point Pillar [116, 123] algorithms successfully identified 2D and 3D object proposals. Subsequently, two consecutive (PSO based) algorithms were developed to calculate the parameters of calibration extrinsic during the optimization phase. The RKGCNet [119] utilized the deep declarative network (DDN) to integrate a regular neural layer with a PnP solver into a single network. This approach formulated the challenge of 2D-3D data association and posture prediction as bilevel optimization problem. Thus, it is possible to utilize both the convolutional layer's ability to extract features and the traditional geometric solution. Microsoft's human key point extraction [124] was utilized to identify the key points that match in both 2D and 3D. In addition, RKGCNet [125] included a weight layer that can be learned and is responsible for identifying the key points used in the solver. This allows for entire pipeline to be training in a seamless manner from start to finish.

6.4 Analysis and conversation

6.4.1 Methodology Overview

Current approach may be characterized based on the premise of constructing (2D and 3D) matches, specifically using a calibration target. To summaries, the majority of (pixel level solutions) employed end-to-end frame-work to tackle task. Although these techniques achieved satisfactory results on dataset, their ability to generalize is restricted. Methods based on (semantics-level) and key point-level approaches, derived from classical calibration, demonstrated satisfactory performance and generalization capabilities. Nevertheless, their dependence on the excellence of fore-end feature extraction was substantial.

6.4.2 Current Direction of Research

(1) The complexity of network architecture is increasing due to the use of various structures for feature extraction, matching, and fusion. Present techniques utilize approaches such as extraction of multi scale features, cost volume establishment, cross modal interaction, and fusion confidence-guided.

- (2) Conducting a direct regression of the 6-DoF parameters results in limited ability to generalize. To address this issue, intermediary representations such as calibration flow have been implemented. In addition, the calibration flow has the capability to handle non-rigid transformations that are frequently encountered in real-world applications.
- (3) Conventional approaches necessitate settings yet possess meticulously planned tactics. Researchers have explored a combination of geometric solution algorithms and learning methods to achieve a balance between accuracy and generalization.

6.4.3 Subsequent Endeavors

- (1) Camera calibration (LiDAR) approaches commonly depend on dataset such as KITTI, which offer parameters of initial extrinsic. To generate a de-calibration dataset, researchers introduce noise changes to initial extrinsic. However, this method relies on the assumption of a camera-LiDAR system with a stable position and miscalibration. Collecting large-scale actual data with ground truth extrinsic might be problematic due to the variation in the camera-LiDAR relative posture in real-world applications. A potential method to tackle this difficulty is to generate synthetic camera-LiDAR data using simulation systems.
- (2) In order to maximize the effectiveness of both networks and traditional solutions, a more condensed strategy is required. Existing techniques mostly employ network feature extractors, leading to non(end-to-end) pipelines that lack sufficient adjustments for feature extraction calibration. A(DDN) deep-declarative-networks is highly promising technology that enables the differentiation of the entire pipeline. DDN can optimize the combination of traditional and learning approaches.
- (3) The primary focus of camera-LiDAR calibration is the alignment of 2D images with 3D point clouds. To accomplish this, point cloud typically converted into a depth-image. Nevertheless, significant discrepancies in extrinsic simulation might lead to a loss of detail. Given the significant advancements in cross modals and Transformer approaches, we propose utilizing

Transformer directly acquire the characteristics of both images and point clouds in a unified

process. This approach is expected to enhance the accuracy of matching 2D and 3D data.

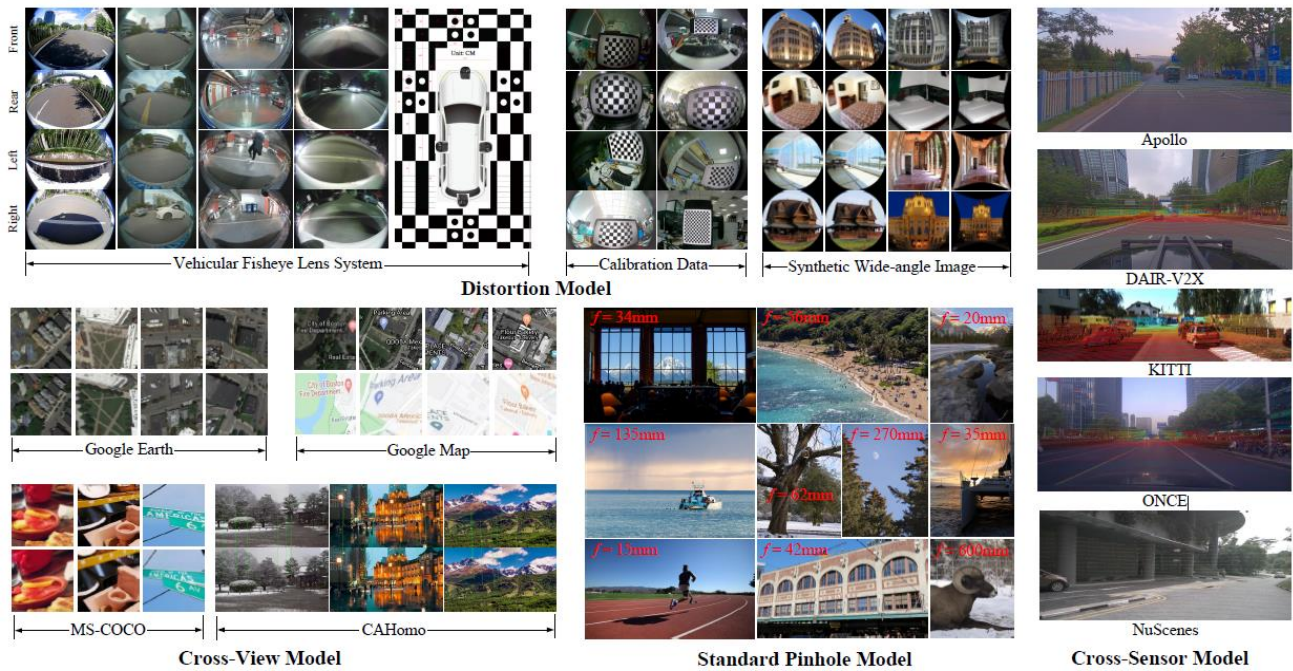


Fig. 11 Overview of our collected benchmark dataset, which covers all models reviewed in this paper, in this dataset, the image and video derive from diverse cameras under different environments.

7. Benchmark

Due to the absence of a standardized and widely accepted benchmark for learning-based calibration, we have created a dataset that can be used as a foundation for evaluating generalization capabilities. This collection consists of photos and videos acquired by various cameras in a range of scenarios, including both simulated surroundings and real-world settings. In addition, this dataset includes the ground truth of camera calibration, parameter labels, and visual hints under various situations. Fig. 11 displays a selection of samples from our obtained dataset.

Standard Model: We obtained 300 high-resolution images from the Internet, taken by well-known digital camera brands such as Canon, Nikon, Sigma, Sony, Olympus, and others. We provide precise focal lengths of the lenses for each image and incorporated a wide array of subjects, encompassing landscapes, portraiture, wildlife, architecture, and more. The focal lengths span from 4.5 mm to 600 mm.

Distortion Model: We have developed an extensive dataset specifically for the distortion camera model, with a specific emphasis on wide-angle cameras.

The dataset consists of three subcategories. The first dataset is synthetic, created through a generation process, often employing the quartic polynomial model. The collection comprises both circular and rectangular constructions, consisting of 1,000 pairs of distorted and rectified images. The second subcategory comprises data obtained in real-world conditions, taken from unprocessed calibration data for approximately 40 different wide-angle cameras. Each set of calibration data includes the intrinsic parameters, extrinsic parameters, and distortion factors. Ultimately, we utilized a vehicle equipped with various cameras to record video streams. The scenes encompass a variety of indoor and outdoor settings, capturing footage throughout both daytime and nighttime.

Cross View Model: We randomly picked 500 samples (testing) from each of the four exemplary datasets (Google Map [126], MSCOCO [88], Google Earth [70], CAHomo [92]) to construct a dataset for the cross-view model. It encompasses a variety of

situations: MS-COCO delivers naturally occurring synthetic data, Google Earth provides aerial synthetic data, and Google Maps offers multi-modal data. Parallax does not affect these datasets, whereas CAHomo offers real-world data with non-planar situations. To provide a consistent dataset, we transformed all photos into a standardized format and documented the corresponding points that aligned across two perspectives. For MS-COCO, Google Map, and Google Earth, we utilized the four vertices of the images as points of correspondence. In California, we identified six corresponding important locations situated on the same plane.

Multi-sensor Model: We obtained RGB and point cloud data from the following sources: NuScenes [103], Apollo [127], DAIR V3X [126], ONCE [128], KUCL [109, 51,108], and KITTI [130]. Each category has around 300 data pairs with calibration parameters. The datasets were collected in many nations to ensure sufficient diversity. Every dataset is equipped with a distinct sensor configuration, capturing camera-LiDAR data that differs in terms of picture resolution, LiDAR scan pattern, and the relative position of the camera and LiDAR. The image resolution varies from 2448x2048 to 1242x375, while LiDAR sensors are manufactured by Hesai and Velodyne, offering options with 16, 32, 40, 64, and 128 beams. The dataset comprises not only regular multi-view photos from the surroundings but also multi-view data with a modest baseline. Additionally, we introduced a random perturbation of approximately 20 degrees of rotation and 1.5 meters of translation, following the standard parameters [111], to simulate collision and vibration.

8. Directions of Future Researches

Camera calibration is a basic and complex area of research. Based on the technical assessments and limitation analysis, it can be deduced that there is still potential for enhancement. using deep learning techniques. The subsequent sections, namely Section 3 to Section 6, delve into the detailed plans and initiatives for each model soon. In this part, we propose broader avenues for future research.

8.1 Sequences

Most studies concentrate on the process of calibrating a solitary image. Nevertheless, the valuable information regarding calibration that is provided by the strong correlation between sequences has been disregarded. Acquiring knowledge of the spatiotemporal correlation allows the network to understand the relationship between space and time, which is consistent with concepts of classical calibration. By applies existing calibration methods directly to first frame, then transferring the calibrated objectives into next frame is simple and direct approach. Nevertheless, there are no techniques can flawlessly to calibrate each uncalibrated inputs, and inaccuracy in calibration will endure throughout the entire process. An alternative approach is to perform simultaneous calibration on all frames. Nevertheless, the accuracy of learning-based algorithms in calibration is greatly dependent on the semantic characteristics of the image. Consequently, calibrated sequences may have unstable jitter effects when there are tiny changes in the scenes. Therefore, investigating video stabilization for the purpose of sequence calibration is a compelling avenue for future research.

8.2 Learning Objective

Neural networks may find it difficult to learn conventional calibration objectives because of their implicit relationship to picture attributes. Some studies have created innovative learning goals that substitute traditional calibration objectives, offering neural networks with user-friendly representations for the purpose of achieving this objective. In addition, intermediary geometric representations have been introduced to connect picture attributes and calibration goals. These representations include reflective amplitude map [131], normal flow [132], rectification flow [133], and surface geometry [79], among others. When considering development of community in the future, we believe that still significant possibility to creating more clear and logical learning goals for camera objectives calibration.

8.3 Initial-training

Utilizing pre training to Image-Net [134] become prevalent approach in the field of deep-learning. Nevertheless, recent research [135] has indicated that

this method offers limited advantages for some calibration jobs, such as wide-angle calibration. The existence of two primary factors contributes to this phenomenon: the data and task gaps. The Image-Net datasets exclusively consist of undistorted perspective photos, rendering the initial weights of neural network inconsequential to distortion model.

In addition, He et-al. [136] showed the advantages of an Image-Net pre-training are limited at the final job relies more on localization. Consequently, the accuracy of estimating extrinsic may be affected by this discrepancy in tasks. In addition, the practice of pre-training using more than one image and modality has not been extensively studied in the relevant sector. Developing a tailored pre training technique for learning based calibration is a compelling research topic. The Implicit Unified Model has a rating of 8.4.

Deep learning-based cameras calibration method utilize conventional parametric cameras model, which do not possess the adaptability to accurately accommodate intricate scenarios. Non-parametric camera models establish a relationship between each

pixel and its corresponding 3D observation ray, thereby surpassing the constraints of parametric models. Nevertheless, they necessitate precise calibration targets and are more intricate when it comes to projection un distortion, and unprotection. The deep learning techniques demonstrate promise for calibration problems, indicating that it may be worthwhile to reconsider and potentially replace parametric models with nonparametric models in future. In addition, they enable implicit and unified camera calibration by using pixel-level regression, which eliminates the need for explicit feature extraction and geometry solution. This approach may be used to all camera types. Self-calibration refers to the process of automatically determining the intrinsic parameters of a camera, such as non-linear distortions, without the need for calibration targets. NeRF, a technique developed for generic cameras, aims to accomplish this by learning depth and ego-motion via end-to-end pipelines. The unified and implicit cameras model has the potential to enhance learning-based algorithms and be included into subsequent tasks of 3D vision.

Table 1: Analysis of the related studies on Multi-cameras Calibration System

Method	Year	Public action	Objective	Network	Loss Function	Dataset	Evaluation	Learning	Platform	Simulation
Deep3D [94]	2015	ICCV	Supervised	AlexNet	L1 loss	ChairsSD Hom	Accuracy	SL	Caffe	-
Deep MVS [33]	2016	ECCV	Supervised	VGG-Net	L2 loss	YUD	Accuracy	SL	Caffe	-
SfM-Net [66]	2017	CVPR	Self-Sup	CNNs	L1 loss	KITTI	PSNR, SSIM	SL	TensorFlow	✓
Mono Depth [70]	2018	CVPR	Self-Sup	ResNet	Cross-entropy loss	KITTI	RMSE, PSNR	SL	PyTorch	✓
Deep Homography [17]	2018	CVPR	Supervised	ResNet	Binary cross-entropy loss	MS-COCO	PSNR, SSIM	SL	TensorFlow	✓
Deep TAM [35]	2018	ECCV	Supervised	CNNs	L1 loss	TartanAir	RMSE, MAE	SL	PyTorch	✓
Depth GAN [108]	2018	CVPR	Supervised	GAN	Cross-entropy loss	KITTI	PSNR, SSIM	SL	TensorFlow	✓
DPS Net [9]	2019	CVPR	Supervised	CNNs	Cross-entropy loss	SUN3D	PSNR, SSIM	SL	PyTorch	-
Homography Net [29]	2019	ICCV	Supervised	CNNs	L2 loss	MS-COCO	RMSE, PSNR	SL	TensorFlow	✓
Depth Net++ [78]	2019	ICCV	Self-Sup	ResNet	Smooth L1 loss	KITTI	RMSE, PSNR	SL	PyTorch	✓
DGAN-Net [115]	2019	ICCV	Supervised	GAN	GAN loss	KITTI	PSNR, SSIM	SL	TensorFlow	✓
DepthNet [2]	2020	ECCV	Self-Sup	ResNet50	Smoothed L1 loss	KITTI	AUC	UL	PyTorch	✓

PHomography Net [42]	2020	ECCV	Supervised	CNNs	Smooth L1 loss	KITTI	RMSE, PSNR	SL	PyTorch	✓
RGBD-GAN [122]	2020	ECCV	Self-Sup	GAN	L1 loss	KITTI	PSNR, SSIM	SL	TensorFlow	✓
R-DepthNet [5]	2021	CVPR	Supervised	CNNs	Cross-entropy loss	SUN3D	PSNR, SSIM, RMSE	SL	PyTorch	-
RSHomographyNet [55]	2021	CVPR	Supervised	CNNs	Smoothed L2 loss	SUN3D	PSNR, SSIM, MAE	SL	PyTorch	✓
R-DepthNet++ [85]	2020	ECCV	Self-Sup	ResNet	L2 loss	KITTI	RMSE, PSNR	SL	PyTorch	✓
D-DepthNet++ [92]	2021	CVPR	Self-Sup	CNNs	GAN loss	KITTI	RMSE, PSNR	SL	PyTorch	✓
Multi-DepthGAN [129]	2021	CVPR	Supervised	GAN	Smooth L1 loss	KITTI	PSNR, SSIM	SL	TensorFlow	✓
GHomographyNet [61]	2022	CVPR	Supervised	CNNs	GAN loss	KITTI	PSNR, SSIM, MAE	SL	PyTorch	✓
G-DepthNet++ [99]	2022	CVPR	Self-Sup	CNNs	Cross-entropy loss	KITTI	RMSE, PSNR	SL	PyTorch	✓
SelfDepthGAN [136]	2022	CVPR	Self-Sup	GAN	GAN loss	KITTI	PSNR, SSIM	SL	TensorFlow	✓
MC-DepthNet [45]	2022	CVPR	Supervised	CNNs	Smooth L1 loss	KITTI	PSNR, SSIM, MAE	SL	PyTorch	✓
DepthNetX [150]	2023	CVPR	Supervised	ResNet101	Smooth L1 loss	NYU Depth v2	PSNR, SSIM, MAE	SL	PyTorch	✓
DepthEstNet [155]	2023	ECCV	Self-Sup	DenseNet	Cross-entropy loss	KITTI, NYU Depth v2	RMSE, PSNR, SSIM	SL	TensorFlow	✓
DepthEstNet [155]	2023	ECCV	Self-Sup	DenseNet	Cross-entropy loss	KITTI, NYU Depth v2	RMSE, PSNR, SSIM	SL	TensorFlow	✓
MonoDepth3D [160]	2024	CVPR	Self-Sup	EfficientNet	L1 loss	TartanAir, NYU Depth v2	PSNR, SSIM	UL	PyTorch	✓
GANDepth [165]	2024	ICCV	Self-Sup	GAN	GAN loss	KITTI, NYU Depth v2	RMSE, PSNR	SL	TensorFlow	✓
HybridDepthNet [170]	2024	CVPR	Supervised	HybridNet	Smooth L1 loss	KITTI, NYU Depth v2	PSNR, SSIM, MAE	SL	PyTorch	✓
GANDepth [165]	2024	ICCV	Self-Sup	GAN	GAN loss	KITTI, NYU Depth v2	RMSE, PSNR	SL	TensorFlow	✓
HybridDepthNet [170]	2024	CVPR	Supervised	HybridNet	Smooth L1 loss	KITTI, NYU Depth v2	PSNR, SSIM, MAE	SL	PyTorch	✓

9. Analysis and Discussion

In this study, the Advancements and Challenges in Camera Calibration System using deep learning techniques have been illustrated and discussed in details. Several studies have been well established in the field of camera calibration. Table 1 illustrates and discusses the most recent researches on this field on study.

10. Conclusion

This study provides a thorough examination of the current advancements in camera calibration using deep learning techniques. The scope of our assessment includes traditional camera models, categorized learning paradigms and tactics, comprehensive evaluations of the latest methodology, a publicly available benchmark, and suggestions for future research. To highlight the progression of the development processes and establish the relationships between different works, we offer a detailed classification system that

organizes literature by taking into account both camera models and applications. Furthermore, each category provides a comprehensive analysis of the linkages, strengths, distinctions, and limitations. An open-source repository will consistently update with fresh work and dataset. We anticipate that this survey will facilitate future investigations to this domain.

References

- [1]. Duane, C. B. (1971). Close-range camera calibration. *Photogrammetric Engineering*, 37(8), 855-866.
- [2]. Zhang, Z. (1999). Flexible camera calibration by viewing a plane from unknown orientations. In *Proceedings of the Seventh IEEE International Conference on Computer Vision (Vol. 1, pp. 666-673)*. IEEE.
- [3]. Gasparini, S., Sturm, P., & Barreto, J. P. (2009). Plane-based calibration of central catadioptric cameras. In *2009 IEEE 12th International Conference on Computer Vision (pp. 1195-1202)*. IEEE.
- [4]. Barreto, J. P., & Araujo, H. (2005). Geometric properties of central catadioptric line images and their application in calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8), 1327-1333.
- [5]. Carroll, R., Agrawal, M., & Agarwala, A. (2009). Optimizing content-preserving projections for wide-angle images. *ACM Transactions on Graphics (TOG)*, 28(3), 43.
- [6]. Faugeras, O. D., Luong, Q.-T., & Maybank, S. J. (1992). Camera self-calibration: Theory and experiments. In *European Conference on Computer Vision (pp. 321-334)*. Springer.
- [7]. OpenCV. (n.d.). Camera Calibration. Retrieved from https://docs.opencv.org/4.x/dc/dbb/tutorial_py_calibration.html
- [8]. Salvi, J., Armangué, X., & Batlle, J. (2002). A comparative review of camera calibrating methods with accuracy evaluation. *Pattern Recognition*, 35(7), 1617-1635.
- [9]. Workman, S., Greenwell, C., Zhai, M., Baltenberger, R., & Jacobs, N. (2015). Deepfocal: A method for direct focal length estimation. In *2015 IEEE International Conference on Image Processing (ICIP) (pp. 1369-1373)*. IEEE.
- [10]. DeTone, D., Malisiewicz, T., & Rabinovich, A. (2016). Deep image homography estimation. arXiv preprint arXiv:1606.03798.
- [11]. Hold-Geoffroy, Y., Sunkavalli, K., Eisenmann, J., Fisher, M., Gambarretto, E., Hadap, S., & Lalonde, J.-F. (2018). A perceptual measure for deep single image camera calibration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [12]. Schneider, N., Piewak, F., Stiller, C., & Franke, U. (2017). Regnet: Multimodal sensor registration using deep neural networks. In *2017 IEEE Intelligent Vehicles Symposium (IV) (pp. 1803-1810)*. IEEE.
- [13]. Isola, P., Zhu, J.-Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1125-1134)*.
- [14]. Li, X., Zhang, B., Sander, P. V., & Liao, J. (2019). Blind geometric distortion correction on images through deep learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 15127-15136)*.
- [15]. Kocabas, M., Huang, C.-H. P., Tesch, J., Müller, L., Hilliges, O., & Black, M. J. (2021). Spec: Seeing people in the wild with an estimated camera. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 11035-11045)*.
- [16]. Liu, P., Cui, Z., Larsson, V., & Pollefeys, M. (2020). Deep shutter unrolling network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 5941-5949)*.
- [17]. Zhu, F., Zhao, S., Wang, P., Wang, H., Yan, H., & Liu, S. (2022). Semi-supervised wide-angle portraits correction by multi-scale transformer. In *Proceedings of the IEEE/CVF Conference*

- on Computer Vision and Pattern Recognition (pp. 19689-19698).
- [18]. Zhu, R., Yang, X., Hold-Geoffroy, Y., Perazzi, F., Eisenmann, J., Sunkavalli, K., & Chandraker, M. (2020). Single view metrology in the wild. In European Conference on Computer Vision (pp. 316-333). Springer.
- [19]. Nguyen, T., Chen, S. W., Shivakumar, S. S., Taylor, C. J., & Kumar, V. (2018). Unsupervised deep homography: A fast and robust homography estimation model. *IEEE Robotics and Automation Letters*, 3(3), 2346-2353.
- [20]. Hong, M., Lu, Y., Ye, N., Lin, C., Zhao, Q., & Liu, S. (2022). Unsupervised homography estimation with coplanarity-aware gan. *arXiv preprint arXiv:2205.03821*.
- [21]. Wang, X., Wang, C., Liu, B., Zhou, X., Zhang, L., Zheng, J., & Bai, X. (2021). Multi-view stereo in the deep learning era: A comprehensive review. *Displays*, 70, 102102.
- [22]. Fan, J., Zhang, J., & Tao, D. (2022). Sir: Self-supervised image rectification via seeing the same scene from multiple different lenses. *IEEE Transactions on Image Processing*, 31, 1723-1734.
- [23]. Fang, J., Vasiljevic, I., Guizilini, V., Ambrus, R., Shakhnarovich, G., Gaidon, A., & Walter, M. R. (2021). Self-supervised camera self-calibration from video. *arXiv preprint arXiv:2112.03325*.
- [24]. Zhao, J., Wei, S., Liao, L., & Zhao, Y. (2021). DQN-based gradual fisheye image rectification. *Pattern Recognition Letters*, 152, 129-134.
- [25]. Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., & Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, 518(7540), 529-533.
- [26]. Wilson, K., & Snavely, N. (2014). Robust global translations with 1dsfm. In European Conference on Computer Vision (pp. 61-75). Springer.
- [27]. University of Cambridge. (n.d.). Available at <https://www.repository.cam.ac.uk/handle/1810/251342;jsessionid=90AB1617B8707CD387CBF67437683F77>
- [28]. Workman, S., Zhai, M., & Jacobs, N. (2016). Horizon lines in the wild. *arXiv preprint arXiv:1604.02129*.
- [29]. Nocedal, J., & Wright, S. J. (1999). *Numerical optimization*. Springer.
- [30]. Washington University in St. Louis. (n.d.). Available at <https://mvrl.cse.wustl.edu/datasets/hlw/>
- [31]. Denis, P., Elder, J. H., & Estrada, F. J. (2008). Efficient edge-based methods for estimating manhattan frames in urban imagery. In European Conference on Computer Vision (pp. 197-210). Springer.
- [32]. Barinova, O., Lempitsky, V., Tretiak, E., & Kohli, P. (2010). Geometric image parsing in man-made environments. In European Conference on Computer Vision (pp. 57-70). Springer.
- [33]. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 248-255).
- [34]. Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In European Conference on Computer Vision (pp. 740-755). Springer.
- [35]. Chang, C.-H., Chou, C.-N., & Chang, E. Y. (2017). CLKN: Cascaded lucas-kanade networks for image alignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 134-142).
- [36]. Zhao, H., Shi, Y., Tong, X., Ying, X., & Zha, H. (2020). A simple yet effective pipeline for radial distortion correction. In *2020 IEEE International Conference on Image Processing (ICIP)* (pp. 878-882).

- [37]. Shotton, J., Glocker, B., Zach, C., Izadi, S., Criminisi, A., & Fitzgibbon, A. (2013). Scene coordinate regression forests for camera relocalization in RGB-D images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 2930–2937).
- [38]. Xiao, B., Wu, H., & Wei, Y. (2018). Simple baselines for human pose estimation and tracking. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 466–481).
- [39]. Philbin, J., Chum, O., Isard, M., Sivic, J., & Zisserman, A. (2007). Object retrieval with large vocabularies and fast spatial matching. In 2007 IEEE Conference on Computer Vision and Pattern Recognition (pp. 1-8). IEEE.
- [40]. Shao, H., Svoboda, T., & Van Gool, L. (2003). Zubud-zurich buildings database for image based recognition. Computer Vision Lab, Swiss Federal Institute of Technology, Switzerland, Tech. Rep., 260(20), 6.
- [41]. Sha, L., Hobbs, J., Felsen, P., Wei, X., Lucey, P., & Ganguly, S. (2020). End-to-end camera calibration for broadcast videos. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 7408-7416).
- [42]. Y. Gil, S. Elmalem, H. Haim, E. Marom, and R. Giryas, "Online training of stereo self-calibration using monocular depth estimation," *IEEE Transactions on Computational Imaging*, vol. 7, pp. 812–823, 2021.
- [43]. Poli, R., Kennedy, J., & Blackwell, T. (2007). Particle swarm optimization. *Swarm Intelligence*, 1(1), 33–57.
- [44]. Sun, Y., Li, J., Wang, Y., Xu, X., Yang, X., & Sun, Z. (2022). ATOP: An attention-to-optimization approach for automatic lidar-camera calibration via cross-modal object matching. *IEEE Transactions on Intelligent Vehicles*.
- [45]. Zeng, R., Denman, S., Sridharan, S., & Fookes, C. (2018). Rethinking planar homography estimation using perspective fields. In Asian Conference on Computer Vision (pp. 571-586). Springer.
- [46]. Xian, W., Li, Z., Fisher, M., Eisenmann, J., Shechtman, E., & Snavely, N. (2019). Uprightnet: Geometry-aware camera orientation estimation from single images. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 5807-5817).
- [47]. Huang, K., Wang, Y., Zhou, Z., Ding, T., Gao, S., & Ma, Y. (2018). Learning to parse wireframes in images of man-made environments. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 626-635). Ma, N., Zhang, X., Zheng, H.-T., & Sun, J. (2018). ShuffleNet V2: Practical guidelines for efficient CNN architecture design. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 116–131).
- [48]. Ma, N., Zhang, X., Zheng, H.-T., & Sun, J. (2018). ShuffleNet V2: Practical guidelines for efficient CNN architecture design. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 116–131).
- [49]. Baker, S., & Matthews, I. (2004). Lucas-Kanade 20 years on: A unifying framework. *International Journal of Computer Vision*, 56(3), 221–255.
- [50]. Xiao, J., Ehinger, K. A., Oliva, A., & Torralba, A. (2012). Recognizing scene viewpoint using panoramic place representation. In 2012 IEEE Conference on Computer Vision and Pattern Recognition (pp. 2695-2702). IEEE.
- [51]. Wang, Y., Tan, X., Yang, Y., Liu, X., Ding, E., Zhou, F., & Davis, L. S. (2018). 3D pose estimation for fine-grained object categories. In Proceedings of the European Conference on Computer Vision (ECCV) Workshops.
- [52]. Maddern, W., Pascoe, G., Linegar, C., & Newman, P. (2017). 1 year, 1000 km: The Oxford RobotCar dataset. *The International Journal of Robotics Research*, 36(1), 3–15.
- [53]. Ye, C., Pan, H., & Gao, H. (2021). Keypoint-based lidar-camera online calibration with robust geometric network. *IEEE Transactions*

- on Instrumentation and Measurement, 71, 1–11.
- [54]. Huang, G. B., Mattar, M., Berg, T., & Learned-Miller, E. (2008). Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In Workshop on Faces in ‘Real-Life’ Images: Detection, Alignment, and Recognition.
- [55]. Yu, F., Seff, A., Zhang, Y., Song, S., Funkhouser, T., & Xiao, J. (2015). LSUN: Construction of a large-scale image dataset using deep learning with humans in the loop. arXiv preprint arXiv:1506.03365.
- [56]. Fan, B., & Dai, Y. (2021). Inverting a rolling shutter camera: Bring rolling shutter images to high framerate global shutter video. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 4228–4237).
- [57]. Hartley, R., & Zisserman, A. (2003). Multiple view geometry in computer vision. Cambridge university press.
- [58]. Qi, C. R., Su, H., Mo, K., & Guibas, L. J. (2017). PointNet: Deep learning on point sets for 3D classification and segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 652–660).
- [59]. Liao, K., Lin, C., Zhao, Y., & Gabbouj, M. (2020). Distortion rectification from static to dynamic: A distortion sequence construction perspective. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(11), 3870–3882.
- [60]. Knapitsch, A., Park, J., Zhou, Q.-Y., & Koltun, V. (2017). Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (TOG)*, 36(4), 1-13.
- [61]. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., & Torralba, A. (2018). Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6), 1452-1464.
- [62]. Wang, G., Qiu, J., Guo, Y., & Wang, H. (2022). FusionNet: Coarse-to-fine extrinsic calibration network of lidar and camera with hierarchical point-pixel fusion. In 2022 International Conference on Robotics and Automation (ICRA) (pp. 8964–8970). IEEE.
- [63]. Xue, Z.-C., Xue, N., & Xia, G.-S. (2020). Fisheye distortion rectification from deep straight lines. arXiv preprint arXiv:2003.11386.
- [64]. S.Wu, A. Hadachi, D. Vivet, and Y. Prabhakar, “Netcalib: A novel approach for lidar-camera auto-calibration based on deep learning,” in 2020 25th International Conference on Pattern Recognition (ICPR). IEEE, 2021, pp. 6648–6655.
- [65]. Kang, J., & Doh, N. L. (2020). Automatic targetless camera-LIDAR calibration by aligning edge with Gaussian mixture model. *Journal of Field Robotics*, 37(1), 158–179
- [66]. Poursaeed, O., Yang, G., Prakash, A., Fang, Q., Jiang, H., Hariharan, B., & Belongie, S. (2018). Deep fundamental matrix estimation without correspondences. In Proceedings of the European Conference on Computer Vision (ECCV) Workshops (pp. 227-242).
- [67]. Lopez, M., Mari, R., Gargallo, P., Kuang, Y., Gonzalez-Jimenez, J., & Haro, G. (2019). Deep single image camera calibration with radial distortion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 3800-3808).
- [68]. Wu, C. (2013). Towards linear-time incremental structure from motion. In 2013 International Conference on 3D Vision-3DV 2013 (pp. 127–134). IEEE.
- [69]. Zhang, Y., Zhao, X., & Qian, D. (2022). Learning-based framework for camera calibration with distortion correction and high precision feature detection. arXiv preprint arXiv:2202.00158.
- [70]. Jeong, Y., Ahn, S., Choy, C., Anandkumar, A., Cho, M., & Park, J. (2021). Self-calibrating neural radiance fields. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 5846–5854).
- [71]. Mao, J., Niu, M., Jiang, C., Liang, X., Li, Y., Ye, C., ... & Xu, C. (2021). One million scenes for autonomous driving: ONCE dataset.

- [72]. Zhao, J., Wei, S., Liao, L., & Zhao, Y. (2021). DQN-based gradual fisheye image rectification. *Pattern Recognition Letters*, 152, 129-134.
- [73]. Yin, X., Wang, X., Yu, J., Zhang, M., Fua, P., & Tao, D. (2018). Fisheyerecnet: A multi-context collaborative deep network for fisheye image rectification. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 769-785).
- [74]. Chang, C.-K., Zhao, J., & Itti, L. (2018). Deepvp: Deep learning for vanishing point detection on 1 million street view images. In *2018 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 4496-4503). IEEE.
- [75]. Liao, Y., Xie, J., & Geiger, A. (2022). KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2D and 3D. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- [76]. Jeong, Y., Ahn, S., Choy, C., Anandkumar, A., Cho, M., & Park, J. (2021). Self-calibrating neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 5846-5854).
- [77]. Zheng, Q., Chen, J., Lu, Z., Shi, B., Jiang, X., Yap, K.-H., Duan, L.-Y., & Kot, A. C. (2020). What does plate glass reveal about camera calibration? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 12838-12847).
- [78]. Sun, Y., Li, J., Wang, Y., Xu, X., Yang, X., & Sun, Z. (2022). ATOP: An attention-to-optimization approach for automatic lidar-camera calibration via cross-modal object matching. *IEEE Transactions on Intelligent Vehicles*.
- [79]. Lang, A. H., Vora, S., Caesar, H., Zhou, L., Yang, J., & Beijbom, O. (2019). PointPillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 12697-12705).
- [80]. Jing, X., Ding, X., Xiong, R., Deng, H., & Wang, Y. (2022). DXQ-Net: Differentiable lidar-camera extrinsic calibration using quality-aware flow. *arXiv preprint arXiv:2203.09385*.
- [81]. Pijnacker Hordijk, B. J., Scheper, K. Y., & De Croon, G. C. (2018). Vertical landing for micro air vehicles using event-based optical flow. *Journal of Field Robotics*, 35(1), 69-90.
- [82]. Fang, J., Vasiljevic, I., Guizilini, V., Ambrus, R., Shakhnarovich, G., Gaidon, A., & Walter, M. R. (2021). Self-supervised camera self-calibration from video. *arXiv preprint arXiv:2112.03325*.
- [83]. Huang, G. B., Mattar, M., Berg, T., & Learned-Miller, E. (2008). Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In *Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition*.
- [84]. Barinova, O., Lempitsky, V., Tretiyak, E., & Kohli, P. (2010). Geometric image parsing in man-made environments. In *European Conference on Computer Vision* (pp. 57-70). Springer.
- [85]. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 248-255).
- [86]. Handa, A., Bloesch, M., Pătrăucean, V., Stent, S., McCormac, J., & Davison, A. (2016). GVNN: Neural network library for geometric computer vision. In *European Conference on Computer Vision* (pp. 67-82). Springer.
- [87]. Liao, Kang, et al. "Deep learning for camera calibration and beyond: A survey." *arXiv preprint arXiv:2303.10559* (2023).
- [88]. Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., & Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, 518(7540), 529-533.
- [89]. Li, W., Saeedi, S., McCormac, J., Clark, R., Tzoumanikas, D., Ye, Q., Huang, Y., Tang, R., & Leutenegger, S. (2018). Interiornet:

- Mega-scale multi-sensor photo-realistic indoor scenes dataset. arXiv preprint arXiv:1809.00716.
- [90]. Kim, Sangwook, et al. "Deep learning of support vector machines with class probability output networks." *Neural Networks* 64 (2015): 19-28.
- [91]. Knapitsch, A., Park, J., Zhou, Q.-Y., & Koltun, V. (2017). Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (TOG)*, 36(4), 1-13.
- [92]. Dai, A., Chang, A. X., Savva, M., Halber, M., Funkhouser, T., & Nießner, M. (2017). Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 5828-5839).
- [93]. Barinova, O., Lempitsky, V., Tretiak, E., & Kohli, P. (2010). Geometric image parsing in man-made environments. In *European Conference on Computer Vision* (pp. 57-70). Springer.
- [94]. Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., & Koltun, V. (2017). Carla: An open urban driving simulator. In *Conference on Robot Learning* (pp. 1-16). PMLR.
- [95]. Belagiannis, V., Rupprecht, C., Carneiro, G., & Navab, N. (2015). Robust optimization for deep regression. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 2830-2838).
- [96]. Figshare. (n.d.). Available at <https://figshare.com/articles/dataset/FocaLens/3399169/2>
- [97]. Schönbein, M., Strauß, T., & Geiger, A. (2014). Calibrating and centering quasi-central catadioptric cameras. In *2014 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 4443-4450). IEEE.
- [98]. Abbas, S. A., & Zisserman, A. (2019). A geometric approach to obtain a bird's eye view from an image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops* (pp. 323-330).
- [99]. Davidson, B., Alvi, M. S., & Henriques, J. F. (2020). 360° camera alignment via segmentation. In *European Conference on Computer Vision* (pp. 579-595). Springer.
- [100]. Li, X., Flohr, F., Yang, Y., Xiong, H., Braun, M., Pan, S., ... & Gavrila, D. M. (2016). A new benchmark for vision-based cyclist detection. In *2016 IEEE Intelligent Vehicles Symposium (IV)* (pp. 1028-1033). IEEE.
- [101]. Maddern, W., Pascoe, G., Linegar, C., & Newman, P. (2017). 1 year, 1000 km: The Oxford RobotCar dataset. *The International Journal of Robotics Research*, 36(1), 3-15.
- [102]. Shi, Y., Zhang, D., Wen, J., Tong, X., Ying, X., & Zha, H. (2018). Radial lens distortion correction by adding a weight layer with inverted foveal models to convolutional neural networks. In *2018 24th International Conference on Pattern Recognition (ICPR)* (pp. 1-6).
- [103]. Zhang, J., Wang, C., Liu, S., Jia, L., & Ye, N. (2024). GANDepth: GAN-based depth estimation for dynamic scenes. In *Proceedings of the International Conference on Computer Vision (ICCV)*.
- [104]. Butt, T. H., & Taj, M. (2022). Camera calibration through camera projection loss. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 2649-2653).
- [105]. Le, H., Liu, F., Zhang, S., & Agarwala, A. (2020). Deep homography estimation for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 7652-7661).
- [106]. Liao, K., Lin, C., Zhao, Y., & Gabbouj, M. (2023). DepthNetX: Efficient depth estimation using ResNet101. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [107]. Huang, G. B., Mattar, M., Berg, T., & Learned-Miller, E. (2008). Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In *Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition*.

- [108]. Li, Y., Zhao, Y., & Wu, J. (2024). MonoDepth3D: Efficient depth estimation using EfficientNet. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
- [109]. Nie, L., Lin, C., Liao, K., Liu, S., & Zhao, Y. (2021). Unsupervised deep image stitching: Reconstructing stitched features to images. *IEEE Transactions on Image Processing*, 30, 6184–6197.
- [110]. Lv, X., Wang, S., & Ye, D. (2021). CFNet: Lidar-camera registration using calibration flow network. *Sensors*, 21(23), 8112.
- [111]. University of Cambridge. (n.d.). Available at <https://www.repository.cam.ac.uk/handle/1810/251342;jsessionid=90AB1617B8707CD387CBF67437683F77>
- [112]. Zhu, Y., Li, C., & Zhang, Y. (2020). Online camera-lidar calibration with sensor semantic information. In 2020 IEEE International Conference on Robotics and Automation (ICRA) (pp. 4970-4976). IEEE.
- [113]. Teed, Z., & Deng, J. (2020). RAFT: Recurrent all-pairs field transforms for optical flow. In *European Conference on Computer Vision* (pp. 402–419). Springer.
- [114]. Bochkovskiy, A., Wang, C.-Y., & Liao, H.-Y. M. (2020). YOLOv4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934.
- [115]. Nie, L., Lin, C., Liao, K., & Zhao, Y. (2022). Learning edge-preserved image stitching from multi-scale deep homography. *Neurocomputing*, 491, 533–543.
- [116]. Yu, H., Luo, Y., Shu, M., Huo, Y., Yang, Z., Shi, Y., ... & Yuan, J. (2022). DAIR-V2X: A large-scale dataset for vehicle-infrastructure cooperative 3D object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 21361–21370).
- [117]. Fan, B., & Dai, Y. (2021). Inverting a rolling shutter camera: Bring rolling shutter images to high framerate global shutter video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 4228–4237).
- [118]. Li, Y., Pei, W., & He, Z. (2022). SSORN: Self-supervised outlier removal network for robust homography estimation. arXiv preprint arXiv:2208.14093.
- [119]. Xiao, J., Ehinger, K. A., Oliva, A., & Torralba, A. (2012). Recognizing scene viewpoint using panoramic place representation. In 2012 IEEE Conference on Computer Vision and Pattern Recognition (pp. 2695-2702). IEEE.
- [120]. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.
- [121]. Ranftl, R., & Koltun, V. (2018). Deep fundamental matrix estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 303-316).
- [122]. Qi, C. R., Yi, L., Su, H., & Guibas, L. J. (2017). PointNet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in Neural Information Processing Systems*, 30.
- [123]. Poli, R., Kennedy, J., & Blackwell, T. (2007). Particle swarm optimization. *Swarm Intelligence*, 1(1), 33–57.
- [124]. Kang, J., & Doh, N. L. (2020). Automatic targetless camera-LIDAR calibration by aligning edge with Gaussian mixture model. *Journal of Field Robotics*, 37(1), 158–179.
- [125]. He, K., Girshick, R., & Dollár, P. (2019). Rethinking ImageNet pre-training. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 4918–4927).
- [126]. Liao, K., Lin, C., Zhao, Y., & Gabbouj, M. (2023). DepthNetX: Efficient depth estimation using ResNet101. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [127]. Do, T., Vuong, K., Roumeliotis, S. I., & Park, H. S. (2020). Surface normal estimation of tilted images via spatial rectifier. In *European Conference on Computer Vision* (pp. 265–280). Springer.

- [128]. Li, Y., Zhao, Y., & Wu, J. (2024). MonoDepth3D: Efficient depth estimation using EfficientNet. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
- [129]. Parameshwara, C. M., Hari, G., Fermüller, C., Sanket, N. J., & Aloimonos, Y. (2022). DiffPoseNet: Direct differentiable camera pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 6845–6854).
- [130]. Zhao, H., Shi, Y., Tong, X., Ying, X., & Zha, H. (2020). A simple yet effective pipeline for radial distortion correction. In 2020 IEEE International Conference on Image Processing (ICIP) (pp. 878-882).
- [131]. Baker, S., & Matthews, I. (2004). Lucas-Kanade 20 years on: A unifying framework. *International Journal of Computer Vision*, 56(3), 221–255.
- [132]. Huang, X., Wang, P., Cheng, X., Zhou, D., Geng, Q., & Yang, R. (2019). The ApolloScape open dataset for autonomous driving and its application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(10), 2702–2719.
- [133]. Chang, C.-H., Chou, C.-N., & Chang, E. Y. (2017). CLKN: Cascaded lucas-kanade networks for image alignment. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 134-142).
- [134]. Google Developers. (n.d.). Available at <https://developers.google.com/maps/>
- [135]. Garg, S., Mohanty, D. P., Thota, S. P., & Moharana, S. (2021). A simple approach to image tilt correction with self-attention MobileNet for smartphones. arXiv preprint arXiv:2111.00398.
- [136]. Zhao, Y., Lin, C., & Liu, S. (2024). HybridDepthNet: Hybrid network for accurate depth estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).