

Deep Learning for Camera Calibration and Beyond: A Survey

Kang Liao, Lang Nie, Shujuan Huang, Chunyu Lin, Jing Zhang, Yao Zhao, *Fellow, IEEE*, Moncef Gabbouj, *Fellow, IEEE*, Dacheng Tao, *Fellow, IEEE*

Abstract— Camera calibration involves estimating camera parameters to infer geometric features from captured sequences, which is crucial for computer vision and robotics. However, conventional calibration is laborious and requires dedicated collection. Recent efforts show that learning-based solutions have the potential to be used in place of the repeatability works of manual calibrations. Among these solutions, various learning strategies, networks, geometric priors, and datasets have been investigated. In this paper, we provide a comprehensive survey of learning-based camera calibration techniques, by analyzing their strengths and limitations. Our main calibration categories include the standard pinhole camera model, distortion camera model, cross-view model, and cross-sensor model, following the research trend and extended applications. As there is no benchmark in this community, we collect a holistic calibration dataset that can serve as a public platform to evaluate the generalization of existing methods. It comprises both synthetic and real-world data, with images and videos captured by different cameras in diverse scenes. Toward the end of this paper, we discuss the challenges and provide further research directions. To our knowledge, this is the first survey for the learning-based camera calibration (spanned 8 years). The summarized methods, datasets, and benchmarks are available and will be regularly updated at <https://github.com/KangLiao929/Awesome-Deep-Camera-Calibration>.

Index Terms— Camera calibration, Deep learning, Computational photography, Multiple view geometry, 3D vision, Robotics.

1 INTRODUCTION

CAMERA calibration is a fundamental and indispensable field in computer vision and it has a long research history [1], [2], [3], [4], tracing back to around 60 years ago [5]. The first step for many vision and robotics tasks is to calibrate the intrinsic (image sensor and distortion parameters) and/or extrinsic (rotation and translation) camera parameters, ranging from computational photography, and multi-view geometry, to 3D reconstruction. In terms of the task type, there are different techniques to calibrate the standard pinhole camera, fisheye lens camera, stereo camera, light field camera, event camera, and LiDAR-camera system, etc. Figure 1 shows the popular calibration objectives, models, and extended applications in camera calibration.

Traditional methods for camera calibration generally depend on hand-crafted features and model assumptions. These methods can be broadly divided into three categories. The most prevalent one involves using a known calibration target (*e.g.*, a checkerboard) as it is deliberately moved in the 3D scene [6], [7], [8]. Then, the camera captures the target from different viewpoints and the checkerboard corners are detected for calculating the camera parameters. However, such a procedure requires cumbersome manual interactions and it cannot achieve automatic calibration “in the wild”.

Kang Liao, Lang Nie, Shujuan Huang, Chunyu Lin (corresponding author), and Yao Zhao are with the Institute of Information Science, Beijing Jiaotong University (BJTU), Beijing 100044, China, and also with the Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing 100044, China (email: kang_liao@bjtu.edu.cn, nielang@bjtu.edu.cn, shujuanhuang@bjtu.edu.cn, cylin@bjtu.edu.cn, yzhao@bjtu.edu.cn)

Moncef Gabbouj is with the Department of Computing Sciences, Tampere University, 33101 Tampere, Finland (e-mail: moncef.gabbouj@tuni.fi)

Jing Zhang and Dacheng Tao are with the School of Computer Science, Faculty of Engineering, The University of Sydney, Australia (e-mail: jing.zhang1@sydney.edu.au; dacheng.tao@gmail.com)

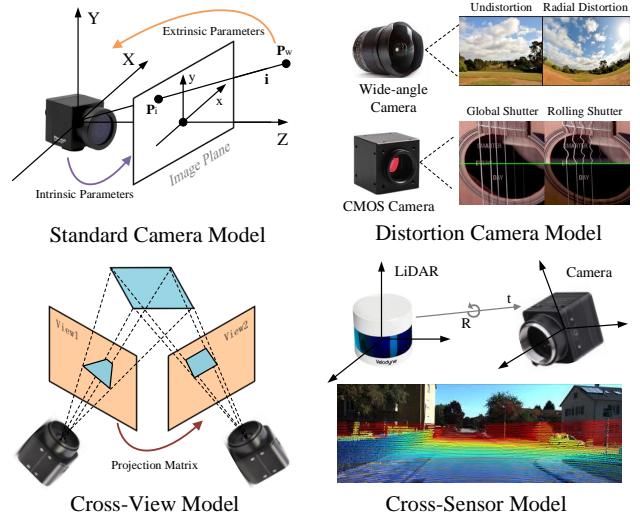


Fig. 1. Popular calibration objectives, models, and extended applications in camera calibration.

To pursue better flexibility, the second category of camera calibration, *i.e.*, the geometric-prior-based calibration has been largely studied [9], [10], [11], [12]. To be specific, the geometric structures are leveraged to model the 3D-2D correspondence in the scene, such as lines and vanishing points. However, this type of method heavily relies on structured man-made scenes containing rich geometric priors, leading to poor performance when applied to general environments. The third category is self-calibration [13], [14], [15]. Such a solution takes a sequence of images as inputs and estimates the camera parameters using multi-view geometry. The accuracy of self-calibration, however, is constrained by the

limits of the feature detectors, which can be influenced by diverse lighting conditions and textures. The above parametric models are based on the physical interpretation of camera geometry. While they are user-friendly, they tend to be tailored for specific camera models and might not offer optimal accuracy. Instead, non-parametric models [16], [17], [18] link each image pixel to its 3D observation ray, eliminating the constraints of parametric models.

Since there are many standard techniques for calibrating cameras in an industry/laboratory implementation [19], [20], this process is usually ignored in recent developments. However, calibrating single and wild images remains challenging, especially when images are collected from websites and unknown camera models. This challenge motivates the researchers to investigate a new paradigm.

Recently, deep learning has brought new inspirations to camera calibration and its applications. Learning-based methods achieve state-of-the-art performances on various tasks with higher efficiency. In particular, diverse deep neural networks (DNNs) have been developed, such as convolutional neural networks (CNNs), generative adversarial networks (GANs), PointNet, and vision transformers (ViTs), of which the high-level semantic features show more powerful representation capability compared with the hand-crafted features. Moreover, diverse learning strategies have been exploited to boost the geometric perception of neural networks. Learning-based methods offer a flexible and end-to-end camera calibration solution, without manual interventions or calibration targets, which sets them apart from traditional methods. Furthermore, some of these methods achieve camera model-free and label-free calibration, showing promising and meaningful applications.

With the rapid increase in the number of learning-based camera calibration methods, it has become increasingly challenging to keep up with new advances. Consequently, there is an urgent need to analyze existing works and foster a community dedicated to this field. Previously, some surveys, *e.g.*, [21], [22], [23] only focused on a specific task/camera in camera calibration or one type of approach. For instance, Salvi et al. [21] reviewed the traditional camera calibration methods in terms of the algorithms. Hughes et al. [22] provided a detailed review for calibrating fisheye cameras with traditional solutions. While Fan et al. [23] discussed both the traditional methods and deep learning methods, their survey only considers calibrating the wide-angle cameras. In addition, due to the few amount of reviewed learning-based methods (around 10 papers), the readers are difficult to picture the development trend of general camera calibration in Fan et al. [23].

In this paper, we provide a comprehensive and in-depth overview of recent advances in learning-based camera calibration, covering over 100 papers. We also discuss potential directions for further improvements and examine various types of cameras and targets. To facilitate future research on different topics, we categorize the current solutions according to calibration objectives and applications. In addition to fundamental parameters such as focal length, rotation, and translation, we also provide detailed reviews for correcting image distortion (radial distortion and rolling shutter distortion), estimating cross-view mapping, calibrating camera-LiDAR systems, and other applications. Such a trend fol-

lows the development of cameras and market demands for virtual reality, autonomous driving, neural rendering, etc.

To our best knowledge, this is the first survey of the learning-based camera calibration and its extended applications, it has the following unique contributions. (1) Our work mainly follows recent advances in deep learning-based camera calibration. In-depth analysis and discussion in various aspects are offered, including publications, network architecture, loss functions, datasets, evaluation metrics, learning strategies, implementation platforms, etc. The detailed information of each literature is listed in Table 1. (2) Despite the calibration algorithm, we comprehensively review the classical camera models and their extended models. In particular, we summarize the redesigned calibration objectives in deep learning since some traditional calibration objectives are verified to be hard to learn by neural networks. (3) We collect a dataset containing images and videos captured by different cameras in different environments, which can serve as a platform to evaluate the generalization of existing methods. (4) We discuss the open challenges of learning-based camera calibration and propose some future directions to provide guidance for further research in this field. (5) An open-source repository is created that provides a taxonomy of all reviewed works and benchmarks. The repository will be updated regularly in <https://github.com/KangLiao929/Awesome-Deep-Camera-Calibration>.

In the following sections, we discuss and analyze various aspects of learning-based camera calibration. The remainder of this paper is organized as follows. In Section 2, we provide the concrete learning paradigms and learning strategies of the learning-based camera calibration. Subsequently, we introduce and discuss the specific methods based on the standard camera model, distortion model, cross-view model, and cross-sensor model in Section 3, Section 4, Section 5, and Section 6, respectively (see Figure 2). The collected benchmark for calibration methods is depicted in Section 7. Finally, we conclude the learning-based camera calibration and suggest the future directions of this community in Section 8.

2 PRELIMINARIES

Deep learning has brought new inspirations to camera calibration, enabling a fully automatic calibration procedure without manual intervention. Here, we first summarize two prevalent paradigms in learning-based camera calibration: regression-based calibration and reconstruction-based calibration. Then, the widely-used learning strategies are reviewed in this research field. The detailed definitions for classical camera models and their corresponding calibration objectives are exhibited in the supplementary material.

2.1 Learning Paradigm

Driven by different architectures of the neural network, the researchers have developed two main paradigms for learning-based camera calibration and its applications.

Regression-based Calibration Given an uncalibrated input, the regression-based calibration first extracts the high-level semantic features using stacked convolutional layers. Then, the fully connected layers aggregate the semantic features and form a vector of the estimated calibration objective.

The regressed parameters are used to conduct subsequent tasks such as distortion rectification, image warping, camera localization, etc. This paradigm is the earliest and has a dominant role in learning-based camera calibration and its applications. All the first works in various objectives, *e.g.*, intrinsics: Deepfocal [24], extrinsic: PoseNet [25], radial distortion: Rong et al. [26], rolling shutter distortion: URS-CNN [27], homography matrix: DHN [28], hybrid parameters: Hold-Geoffroy et al. [29], camera-LiDAR parameters: RegNet [30] have been achieved with this paradigm.

Reconstruction-based Calibration On the other hand, the reconstruction-based calibration paradigm discards the parameter regression and directly learns the pixel-level mapping function between the uncalibrated input and target, inspired by the conditional image-to-image translation [31] and dense visual perception [32], [33]. The reconstructed results are then calculated for the pixel-wise loss with the ground truth. In this regard, most reconstruction-based calibration methods [34], [35], [36], [37] design their network architecture based on the fully convolutional network such as U-Net [38]. Specifically, an encoder-decoder network, with skip connections between the encoder and decoder features at the same spatial resolution, progressively extracts the features from low-level to high-level and effectively integrates multi-scale features. At the last convolutional layer, the learned features are aggregated into the target channel, reconstructing the calibrated result at the pixel level.

In contrast to the regression-based paradigm, the reconstruction-based paradigm does not require the label of diverse camera parameters. Besides, the imbalance loss problem can be eliminated since it only optimizes the photometric loss of calibrated results. Therefore, the reconstruction-based paradigm enables a blind camera calibration without a strong camera model assumption.

2.2 Learning Strategies

In the following, we review the learning-based camera calibration literature regarding different learning strategies.

Supervised Learning Most learning-based camera calibration methods train their networks with the supervised learning strategy, from the classical methods [24], [25], [26], [28], [39], [40] to the state-of-the-art methods [41], [42], [43], [44], [45]. In terms of the learning paradigm, this strategy supervises the network with the ground truth of the camera parameters (regression-based paradigm) or paired data (reconstruction-based paradigm). In general, they synthesize the training dataset from other large-scale datasets, under the random parameter/transformation sampling and camera model simulation. Some recent works [46], [47], [48], [49] establish their training dataset using a real-world setup and label the captured images with manual annotations, thereby fostering advancements in this research domain.

Semi-Supervised Learning Training the network using an annotated dataset under diverse scenarios is an effective learning strategy. However, human annotation can be prone to errors, leading to inconsistent annotation quality or the inclusion of contaminated data. Consequently, increasing the training dataset to improve performance can be challenging due to the complexity and cost of constructing the dataset. To address this challenge, SS-WPC [50] proposes a semi-supervised method for correcting portraits captured

by a wide-angle camera. It employs a surrogate task (segmentation) and a semi-supervised method that utilizes direction and range consistency and regression consistency to leverage both labeled and unlabeled data.

Weakly-Supervised Learning Although significant progress has been made, data labeling for camera calibration is a notorious costly process, and obtaining perfect ground-truth labels is challenging. As a result, it is often preferable to use weak supervision with machine learning methods. Weakly supervised learning refers to the process of building prediction models through learning with inadequate supervision. Zhu et al. [51] present a weakly supervised camera calibration method for single-view metrology in unconstrained environments, where there is only one accessible image of a scene composed of objects of uncertain sizes. This work leverages 2D object annotations from large-scale datasets, where people and buildings are frequently present and serve as useful “reference objects” for determining 3D size.

Unsupervised Learning Unsupervised learning, commonly referred to as unsupervised machine learning, analyzes and groups unlabeled datasets using machine learning algorithms. UDHN [52] is the first work for a cross-view camera model using unsupervised learning, which estimates the homography matrix of a paired image without the projection labels. By reducing a pixel-wise intensity error that does not require ground truth data, UDHN [52] outperforms previous supervised learning techniques. While preserving superior accuracy and robustness to fluctuation in light, the proposed unsupervised algorithm can also achieve faster inference time. Inspired by this work, increasing more methods leverage the unsupervised learning strategy to estimate the homography such as CA-UDHN [53], BaseHomo [54], HomoGAN [55], and Liu et al. [56]. Besides, UnFishCor [57] frees the demands for distortion parameters and designs an unsupervised framework for the wide-angle camera.

Self-supervised Learning Robotics is where the phrase “self-supervised learning” first appears, as training data is automatically categorized by utilizing relationships between various input sensor signals. Compared to supervised learning, self-supervised learning leverages input data itself as the supervision. Many self-supervised techniques are presented to learn visual characteristics from massive amounts of unlabeled photos or videos without the need for time-consuming and expensive human annotations. SSR-Net [58] presents a self-supervised deep homography estimation network, which relaxes the need for ground truth annotations and leverages the invertibility constraints of homography. To be specific, SSR-Net [58] utilizes the homography matrix representation in place of other approaches’ typically-used 4-point parameterization, to apply the invertibility constraints. SIR [59] devises a brand-new self-supervised camera calibration pipeline for wide-angle image rectification, based on the principle that the corrected results of distorted images of the same scene taken with various lenses need to be the same. With self-supervised depth and pose learning as a proxy aim, Fang et al. [60] present to self-calibrate a range of generic camera models from raw video, offering for the first time a calibration evaluation of camera model parameters learned solely via self-supervision.

Reinforcement Learning Instead of aiming to minimize at each stage, reinforcement learning can maximize the

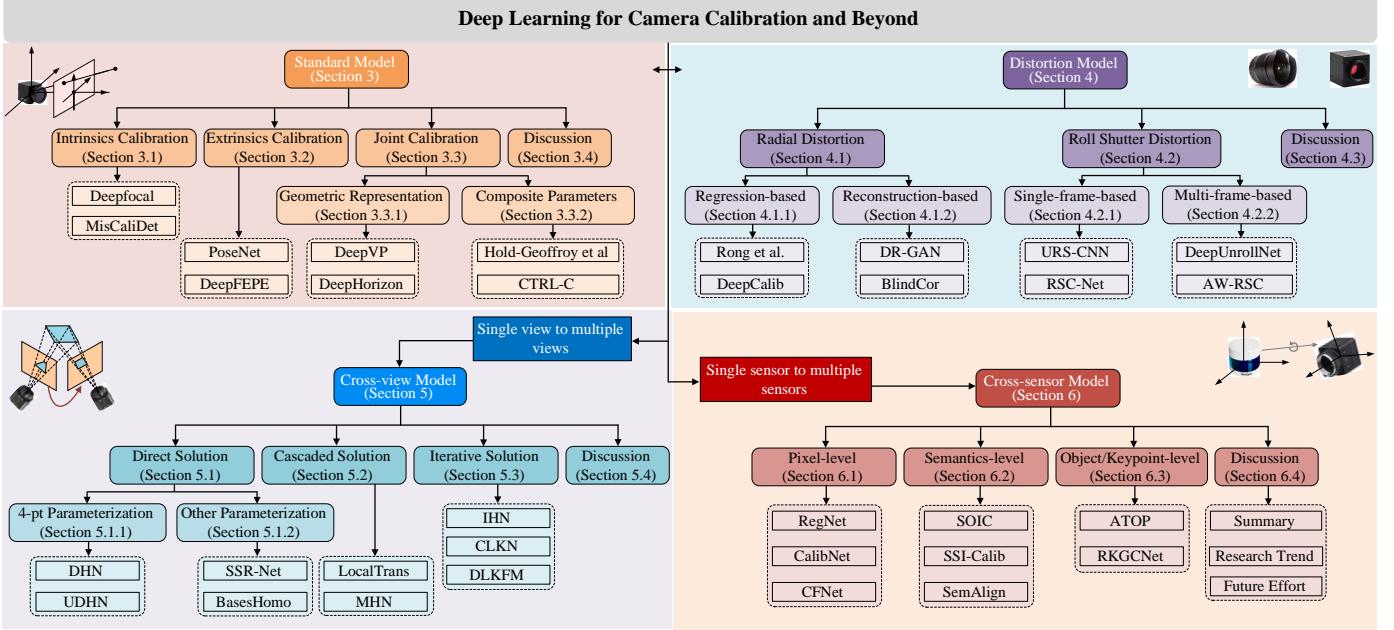


Fig. 2. The structural and hierarchical taxonomy of camera calibration with deep learning. Some classical methods are listed under each category.

cumulative benefits of a learning process as a whole. To date, DQN-RecNet [61] is the first and only work in camera calibration using reinforcement learning. It applies a deep reinforcement learning technique to tackle the fisheye image rectification by a single Markov Decision Process, which is a multi-step gradual calibration scheme. In this situation, the current fisheye image represents the state of the environment. The agent, Deep Q-Network [62], generates an action that should be executed to correct the distorted image.

In the following, we will review the specific methods and literature for learning-based camera calibration. The structural and hierarchical taxonomy is shown in Figure 2.

3 STANDARD MODEL

Generally, for learning-based calibration works, the objectives of the intrinsics calibration contain focal length and optical center, and the objectives of the extrinsic calibration contain the rotation matrix and translation vector.

3.1 Intrinsics Calibration

Deepfocal [24] is a pioneer work in learning-based camera calibration, it aims to estimate the focal length of any image “in the wild”. In detail, Deepfocal considered a simple pinhole camera model and regressed the horizontal field of view using a deep convolutional neural network. Given the width w of an image, the relationship between the horizontal field of view H_θ and focal length f can be described by:

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

Due to component wear, temperature fluctuations, or outside disturbances like collisions, the calibrated parameters of a camera are susceptible to change over time. To this end, MisCalibDet [111] proposed to identify if a camera needs to be recalibrated intrinsically. Compared to the conventional intrinsic parameters such as the focal length and

image center, MisCalibDet presented a new scalar metric, *i.e.*, the average pixel position difference (APPD) to measure the degree of camera miscalibration, which describes the mean value of the pixel position differences over the entire image.

3.2 Extrinsic Calibration

In contrast to intrinsic calibration, extrinsic calibration infers the spatial correspondence of the camera and its located 3D scene. PoseNet [25] first proposed deep convolutional neural networks to regress 6-DoF camera pose in real-time. A pose vector \mathbf{p} was predicted by PoseNet, given by the 3D position \mathbf{x} and orientation represented by quaternion \mathbf{q} of a camera, namely, $\mathbf{p} = [\mathbf{x}, \mathbf{q}]$. For constructing the training dataset, the labels are automatically calculated from a video of the scenario using a structure from motion method [186].

Inspired by PoseNet [25], the following works improved the extrinsic calibration in terms of the intermediate representation, interpretability, data format, learning objective, etc. For example, to optimize the geometric pose objective, DeepFEPE [115] designed an end-to-end keypoint-based framework with learnable modules for detection, feature extraction, matching, and outlier rejection. Such a pipeline imitated the traditional baseline, in which the final performance can be analyzed and improved by the intermediate differentiable module. To bridge the domain gap between the extrinsic objective and image features, recent works proposed to first learn an intermediate representation from the input, such as surface geometry [89], depth map [137], directional probability distribution [151], and normal flow [170], etc. Then, the extrinsic are reasoned by geometric constraints and learned representation. Therefore, the neural networks are gradually guided to perceive the geometry-related features, which are crucial for extrinsic estimation. Considering the privacy concerns and limited storage problem, some recent works compressed the scene and exploited the point-like feature to estimate the extrinsic. For example,

TABLE 1

Details of the learning-based camera calibration and its extended applications from 2015 to 2022, including the method abbreviation, publication, calibration objective, network architecture, loss function, dataset, evaluation metrics, learning strategy, platform, and simulation or not (training data). For the learning strategies, SL, USL, WSL, Semi-SL, SSL, and RL denote supervised learning, unsupervised learning, weakly-supervised learning, semi-supervised learning, self-supervised learning, and reinforcement learning, respectively.

	Method	Publication	Objective	Network	Loss Function	Dataset	Evaluation	Learning	Platform	Simulation	
2015	DeepFocal [24]	ICIP	Standard	AlexNet	L_2 loss	1DSIM [63]	Accuracy	SL	Caffe		
	PoseNet [25]	ICCV	Standard	GoogLeNet	L_2 loss	Cambridge Landmarks [64]	Accuracy	SL	Caffe		
2016	DeepHorizon [65]	BMVC	Standard	GoogLeNet	Huber loss	HLW [66]	Accuracy	SL	Caffe		
	DeepVP [39]	CVPR	Standard	AlexNet	Logistic loss	YUD [67], ECD [68], HLW [66]	Accuracy	SL	Caffe		
	Rong et al. [26]	ACCV	Distortion	AlexNet	Softmax loss	ImageNet [69]	Line length	SL	Caffe	✓	
	DHN [28]	RSSW	Cross-View	VGG	L_2 loss	MS-COCO [70]	MSE	SL	Caffe	✓	
2017	CLKN [71]	CVPR	Cross-View	CNNs	Hinge loss	MS-COCO [70]	MSE	SL	Torch	✓	
	HierarchicalNet [72]	ICCV	Cross-View	VGG	L_2 loss	MS-COCO [70]	MSE	SL	TensorFlow		
	URS-CNN [27]	CVPR	Distortion	CNNs	L_2 loss	Sun [73], Oxford [74], Zubud [75], LFW [76]	PSNR, RMSE	SL	Torch	✓	
	RegNet [30]	IV	Cross-Sensor	CNNs	L_2 loss	KITTI [77]	MAE	SL	Caffe	✓	
2018	Hold-GeoNet [29]	CVPR	Standard	DenseNet	Entropy loss	SUN360 [78]	Human sensitivity	SL	-		
	DeepCalib [40]	CVMP	Distortion	Inception-V3	Logcosh loss	SUN360 [78]	Mean error	SL	TensorFlow		
	FishEyeRecNet [79]	ECCV	Distortion	VGG	L_2 loss	ADE20K [80]	PSNR, SSIM	SL	Caffe	✓	
	Shi et al. [81]	ICPR	Distortion	ResNet	L_2 loss	ImageNet [69]	MSE	SL	PyTorch	✓	
	DeepFm [82]	ECCV	Cross-View	ResNet	L_2 loss	T&T [83], KITTI [77], 1DSIM [63]	F-score, Mean	SL	PyTorch	✓	
	Poursaeed et al. [84]	ECCVW	Cross-View	CNNs	L_1, L_2 loss	KITTI [77]	EPI-ABS, EPI-SQR	SL	-		
	UDHN [52]	RAL	Cross-View	VGG	L_1 loss	MS-COCO [70]	RMSE	USL	TensorFlow	✓	
	PFNet [85]	ACCV	Cross-View	FCN	Smooth L_1 loss	MS-COCO [70]	MAE	SL	TensorFlow	✓	
	CalibNet [86]	IROIS	Cross-Sensor	ResNet	Point cloud distance, L_2 loss	KITTI [77]	Geodesic distance, MAE	SL	TensorFlow	✓	
	Chang et al. [87]	ICRA	Standard	AlexNet	Cross-entropy loss	DeepVP-IM [87]	MSE, Accuracy	SL	Matconvnet		
2019	Lopez et al. [88]	CVPR	Distortion	DenseNet	Bearing loss	SUN360 [78]	MSE	SL	PyTorch		
	UprightNet [89]	ICCV	Standard	U-Net	Geometry loss	InteriorNet [90], ScanNet [91], SUN360 [78]	Mean error	SL	PyTorch		
	Zhuang et al. [92]	IROIS	Distortion	ResNet	L_1 loss	KITTI [77]	Mean error, RMSE	SL	PyTorch	✓	
	SSR-Net [58]	PRL	Cross-View	ResNet	L_2 loss	MS-COCO [70]	MAE	SSL	PyTorch	✓	
	Abbas et al. [93]	ICCVW	Cross-View	CNNs	Softmax loss	CARLA [94]	AUC [95], Mean error	SL	TensorFlow	✓	
	DR-GAN [34]	TCSTV	Distortion	GANs	Perceptual loss	MS-COCO [70]	PSNR, SSIM	SL	TensorFlow	✓	
	STD [96]	TCSTV	Distortion	GANs+CNNs	Perceptual loss	MS-COCO [70]	PSNR, SSIM	SL	TensorFlow	✓	
	Deep360Up [97]	VR	Standard	DenseNet	Log-loss [98]	SUN360 [78]	Mean error	SL	-		
	UnFishCov [57]	JVCIR	Distortion	VGG	L_1 loss	Place2 [99]	PSNR, SSIM	USL	TensorFlow	✓	
	BlindCor [37]	CVPR	Distortion	U-Net	L_2 loss	Place2 [99]	MSE	SL	PyTorch	✓	
	RSC-Net [100]	CVPR	Distortion	ResNet	L_1 loss	KITTI [77]	Mean error	SL	PyTorch	✓	
	Xue et al. [101]	CVPR	Distortion	ResNet	L_2 loss	Wireframes [102], SUNCG [103]	PSNR, SSIM, RPE	SL	PyTorch	✓	
	Zhai et al. [46]	ICCV	Distortion	VGG+U-Net	L_1 loss	Self-constructed+BU-4DFE [104]	Mean error	SL	-		
	NeurIPS [105]	NeurIPS	Standard	CNNs	Binary cross entropy, chamfer- L_2 loss	ScanNet [91], SU3 [106]	Angle accuracy	SL	PyTorch		
2020	Sha et al. [107]	CVPR	Cross-View	U-Net	Cross-entropy loss	World Cup 2014 [108]	IoU	SL	TensorFlow		
	Lee et al. [108]	ECCV	Standard	PointNet + CNNs	Cross-entropy loss	Google Street View [110], HLW [66]	Mean error, AUC [95]	SL	-		
	Mis-ClaiDit [111]	ICRA	Distortion	CNNs	L_2 loss	KITTI [77]	MSE	SL	TensorFlow	✓	
	DeepPTZ [112]	WACV	Distortion	Inception-V3	L_1 loss	SUN360 [78]	Mean error	SL	PyTorch	✓	
	Davidson et al. [113]	CVPR	Cross-View	VGG	Cross-entropy loss	MS-COCO [70], Self-constructed	MAE	SL	TensorFlow	✓	
	CA-UDHN [53]	ECCV	Cross-View	FCN	Dice loss	SUN360 [78]	Accuracy	SL	-		
	DeepFEPE [115]	IROIS	Standard	VGG + PointNet	Triplet loss	KITTI [77], ApolloScape [116]	MSE	USL	PyTorch		
	DDM [35]	TIP	Distortion	GANs	L_1 loss	MS-COCO [70]	Mean error	SL	PyTorch		
	Li et al. [117]	TIP	Distortion	CNNs	Cross-entropy, L_1 loss	Celeba [118]	PSNR, SSIM	TensorFlow	✓		
	PSE-GAN [119]	ICPR	Distortion	GANs	$L_1, WGAN$ loss	Place2 [99]	Cosine distance	SL	-		
	RDC-Net [120]	ICIP	Distortion	ResNet	L_1, L_2 loss	ImageNet [69]	MSE	SL	PyTorch	✓	
	FE-GAN [121]	ICASSP	Distortion	GANs	L_1, GAN loss	Wireframe [102], LSUN [122]	PSNR, SSIM, RMSE	SSL	PyTorch	✓	
	RDFace [123]	CVPR	Distortion	ResNet	L_2 loss	IMDB-Face [124]	Accuracy	SL	-		
	LaRecNet [125]	arXiv	Distortion	ResNet	Cross-entropy, L_2 loss	Wireframes [102], SUNCG [103]	PSNR, SSIM, RPE	SL	PyTorch	✓	
	Baradad et al. [126]	CVPR	Standard	CNNs	L_2 loss	ScanNet [91], NYU [127], SUN360 [78]	Mean error, RMS	SL	PyTorch	✓	
	Zheng et al. [128]	CVPR	Standard	CNNs	L_1 loss	FocalLens [129]	Mean error, PSNR, SSIM	SL	PyTorch	✓	
	Zhu et al. [51]	ECVG	Standard	CNNs + PointNet	L_1 loss	SUN360 [78], MS-COCO [70]	Mean error, Accuracy	WSL	PyTorch	✓	
	DeepUnrollNet [49]	CVPR	Distortion	FCN	L_1 , perceptual, total variation loss	Carla-RS [49], Fastec-RS [49]	PSNR, SSIM, RMSE	SL	PyTorch	✓	
	RGNet [130]	RAL	Cross-Sensor	ResNet	Geodesic distance loss	KITTI [77]	MSE, MSEE, MRR	SL	TensorFlow	✓	
	CalibRCNN [131]	IROIS	Cross-Sensor	RNNs	L_2 , Epipolar geometry loss	KITTI [77]	MAE	SL	TensorFlow	✓	
	SSI-Calib [132]	ICRA	Cross-Sensor	CNNs	L_2 loss	Pascal VOC 2012 [133]	Mean/standard deviation	SL	TensorFlow	✓	
	SOIC [134]	arXiv	Cross-Sensor	ResNet + PointRCNN	Cost function	KITTI [77]	Mean error	SL	PyTorch	✓	
	Net4Calib [135]	ICPR	Cross-Sensor	CNNs	L_1 loss	KITTI [77]	MAE	SL	PyTorch	✓	
	SRHEN [136]	ACM-MM	Cross-View	CNNs	L_2 loss	MS-COCO [70], SUN397 [78]	MACE	SL	-		
2021	StereoCalNet [127]	TGJ	Standard	U-Net	L_1 loss	TAU-Agent [138], KITTI [77]	Mean error	SL	PyTorch	✓	
	CTRL-C [139]	ICCV	Standard	Transformer	Cross-entropy, L_1 loss	Google Street View [110], SUN360 [78]	Mean error, AUC [95]	SL	PyTorch	✓	
	Wakai et al. [140]	ICCVW	Distortion	DenseNet	Smooth L_1 loss	StreetLearn [141]	Mean error, PSNR, SSIM	SL	-		
	OrdianDistortion [142]	TIP	Distortion	CNNs	Smooth L_1 loss	MS-COCO [70], SSIM, MLDL	PSNR, SSIM, MLDL	TensorFlow	✓		
	PolarRecNet [143]	TCSTV	Distortion	VGG + U-Net	L_1, L_2 loss	MS-COCO [70], LMS [144]	PSNR, SSIM, MSE	SL	PyTorch	✓	
	DQN-RecNet [61]	PRL	Distortion	VGG	L_2 loss	Wireframe [102]	PSNR, SSIM, MSE	RL	PyTorch	✓	
	Tan et al. [47]	CVPR	Distortion	U-Net	L_2 loss	Self-constructed	Accuracy	SL	PyTorch	✓	
	DataRecNet [36]	ICCV	Distortion	U-Net	L_1, L_2, GAN loss	Place2 [99]	Place2 [99]	PSNR, SSIM, FID, CW-SSIM	SL	PyTorch	✓
	DLKFM [146]	CVPR	Cross-View	Siamese-Net	Smooth L_1, L_2 loss	MS-COCO [70], Google Earth, Google Map	PSNR, SSIM	SL	PyTorch	✓	
	LocalTrans [147]	ICCV	Cross-View	Transformer	L_1 loss	MS-COCO [70]	MSE, PSNR, SSIM	SL	PyTorch	✓	
	BasesHomeo [144]	ICCV	Cross-View	ResNet	Triplet loss	CA-UDHN [53]	MSE	USL	PyTorch	✓	
	ShuffleHomeoNet [148]	ICIP	Cross-View	ShuffleNet	L_2 loss	MS-COCO [70]	RMSE	SL	TensorFlow	✓	
	DAMG-Homo [144]	TCSTV	Cross-View	CNNs	L_1 loss	MS-COCO [70], TUD [149]	MAE, Accuracy	SL	TensorFlow	✓	
	SA-MobileNet [150]	BMVC	Standard	MobileNet	Cross-entropy loss	SUN360 [78], ADE20K [80], NYU [127]	W-MPJPE, PA-MPJPE	SL	PyTorch	✓	
	SPEC [48]	ICCV	Standard	ResNet	Softargmax- L_2 loss	Self-constructed	Mean and median error	SL	TensorFlow	✓	
	DirectionNet [151]	CVPR	Standard	U-Net	Cosine similarity loss	InteriorNet [90], Matterport3D [152]	Mean and median error	PSNR, SSIM, LPPIs	SL	TensorFlow	✓
	JCD [153]	CVPR	Distortion	FCN	Charbonnier [154], perceptual loss	BS-R3D [153], Fastec-RS [49]	MSE	SL	PyTorch	✓	
	LCCNet [155]	CVPRW	Cross-Sensor	CNNs	Smooth L_1, L_2 loss	KITTI [77]	MAE, MSEE, MRR	SL	PyTorch	✓	
	CFNet [156]	Sensors	Cross-Sensor	FCN	$L_1, Charbonnier$ [154] loss	Carla-RS [49], Fastec-RS [49]	PSNR, SSIM, LPPIs	SL	PyTorch	✓	
	Fan et al. [158]	ICCV	Distortion	U-Net	L_1 , perceptual loss	Carla-RS [49], Fastec-RS [49]	PSNR, SSIM	SL	PyTorch	✓	
	SUN [159]	ICCV	Distortion	DenseNet + ResNet	L_1 , perceptual loss	KITTI [77]	Mean/median rotation errors	SL	PyTorch	✓	
	SemAlign [160]	IROIS	Cross-Sensor	CNNs	Semantic alignment loss						
2022	DVDNet [41]	CVPR	Standard	CNNs	Cross-entropy loss	SU3 [106], ScanNet [91], YUD [67], NYU [127]	Accuracy, AUC [98]	SL	PyTorch	✓	
	Fang et al. [60]	ICRA	Standard	CNNs	L_2 loss	KITTI [77], EuRoC [161], OmniCam [162]	MRE, RMSE	SSL	PyTorch	✓	
	CPL [163]	ICASSP	Standard	Inception-V3	L_1 loss	CARLA [94], CyclistDetection [164]	MAE	SL	TensorFlow	✓	
	IHN [165]	CVPR	Cross-View	Siamese-Net	L_1 loss	MS-COCO [70], Google Earth, Google Map	MACE	SL	PyTorch	✓	
	HomogGAN [55]	CVPR	Cross-View	GANs	Cross-entropy, WGAN loss	CA-UDHN [53]	Mean error	USL	PyTorch	✓	
	SS-WPC [50]	CVPR	Distortion	Transformer	Cross-entropy, L_1 loss	Tan et al. [47]	Accuracy	Semi-SSL	PyTorch	✓	
	AW-RSC [166]	CVPR	Distortion	CNNs	Charbonnier [154], perceptual loss	Self-constructed, FastecRS [49]	PSNR, SSIM, LPPIs	SL	PyTorch	✓	
	EvUnroll [42]	CVPR	Distortion	U-Net	Charbonnier, perceptual, TV loss	Self-constructed, FastecRS [49]	PSNR, SSIM, LPPIs	SL	PyTorch	✓	
	Do et al. [167]	CVPR	Distortion	ResNet	L_2 , Robust angular [168] loss	Self-constructed, 7-SCENES [169]	Median error, Recall	SL	PyTorch	✓	
	DiffPoseNet [170]	CVPR	Standard	CNNs + LSTM	L_2 loss	TartanAir [171], KITTI [77], TUM-RGBD [172]	PEE, AEE [173]	SSL	PyTorch	✓	
	SceneSqueezeNet [174]	CVPR	Standard	Transformer	L_1 loss	RobotCar Scenes [175], Cambridge Landmarks [64]	Mean error, Recall [173]	SL	PyTorch	✓	
	FocalPose [176]	CVPR	Standard	CNNs	L_1 , Huber loss	KITTI [77], TUM-RGBD [172]	Median error, Accuracy	SL	PyTorch	✓	
	DXQ-Net [177]	arXiv	Cross-Sensor	CNNs + RNNs	L_1 , geodesic loss	Pix3D [177], CompCars [178], StanfordCars [178]	MSE	SL	PyTorch	✓	
	SST-Calib [145]	ITSC	Cross-Sensor	CNNs	L_2 loss	KITTI [77]	QAD, AEAD	SL	PyTorch	✓	
	CCS-NET [180]	IROIS	Distortion	U-Net	L_1 loss	MAE, RPE	SL	PyTorch	✓		
	FishFormer [43]	arXiv	Distortion	Transformer	L_2 loss	Place2 [99], Celeba [118]	PSNR, SSIM, FID	SL	PyTorch	✓	
	SIR [59]	TIP	Distortion	ResNet	L_1 loss	ADE20K [80], Wireframes [102], MS-COCO [70]	PSNR, SSIM	SSL	PyTorch	✓	
	ATOP [181]	TIP	Cross-Sensor	CNNs	Cross entropy loss	Self-constructed + KITTI [77]	RRE, RTE	SL	-		
	FusionNet [182]	ICRA	Cross-Sensor	CNNs+PointNet	L_2 loss	KITTI [77]	MAE	SL	PyTorch	✓	
	RGKNet [183]	TIME	Cross-Sensor	CNNs+PointNet	L_1 loss	KITTI [77]	MSE	SL	PyTorch	✓	
	GenCalNet [184]	ECCV	Distortion	DenseNet	L_2 loss	StreetLearn [141], SP360 [185]	MAE, PSNR, SSIM	SL	-		
	Liu et al. [56]	TPAMI	Cross-View	ResNet	Triplet loss	Self-constructed	MSE, Accuracy	USL	PyTorch	✓	

Do et al. [167] trained a network to recognize sparse but significant 3D points, dubbed scene landmarks, by encoding their appearance as implicit features. And the camera pose can be calculated using a robust minimal solver followed by a Levenberg-Marquardt-based nonlinear refinement. SceneSqueezer [174] compressed the scene information from three levels: the database frames are clustered using pairwise co-visibility information, a point selection module prunes each cluster based on estimation performance, and learned quantization further compresses the selected points.

3.3 Joint Intrinsic and Extrinsic Calibration

3.3.1 Geometric Representations

Vanishing Points The intersection of projections of a set of parallel lines in the world leads to a vanishing point. The detection of vanishing points is a fundamental and crucial challenge in 3D vision. In general, vanishing points reveal the direction of 3D lines, allowing the agent to deduce 3D scene information from a single 2D image.

DeepVP [39] is the first learning-based work for detecting the vanishing points given a single image. It reversed the conventional process by scoring the horizon line candidates according to the vanishing points they contain. Chang et al. [87] redesigned this task as a CNN classification problem using an output layer with 225 discrete possible vanishing point locations. For constructing the dataset, the camera view is panned and tilted with step 5° from -35° to 35° in the panorama scene (total 225 images) from a single GPS location. To directly leverage the geometric properties of vanishing points, NeurVPS [105] proposed a canonical conic space and a conic convolution operator that can be implemented as regular convolutions in this space, where the learning model is capable of calculating the global geometric information of vanishing points locally. To overcome the need for a large amount of training data in previous methods, DVDPD [41] incorporated the neural network with two geometric priors: Hough transformation and Gaussian sphere. First, the convolutional features are transformed into a Hough domain, mapping lines to distinct bins. The projection of the Hough bins is then extended to the Gaussian sphere, where lines are transformed into great circles and vanishing points are located at the intersection of these circles. Geometric priors are data-efficient because they eliminate the necessity for learning this information from data, which enables an interpretable learning framework and generalizes better to domains with slightly different data distributions.

Horizon Lines The horizon line is a crucial contextual attribute for various computer vision tasks especially image metrology, computational photography, and 3D scene understanding. The projection of the line at infinity onto any plane that is perpendicular to the local gravity vector determines the location of the horizon line.

Given the FoV, pitch, and roll of a camera, it is straightforward to locate the horizon line in its captured image space. DeepHorizon [65] proposed the first learning-based solution for estimating the horizon line from an image, without requiring any explicit geometric constraints or other cues. To train the network, a new benchmark dataset, Horizon Lines in the Wild (HLW), was constructed, which

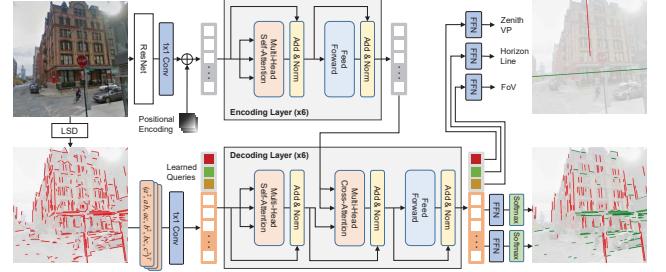


Fig. 3. Overview of CTRL-C. The figure is from [139]. It estimates parameters including the zenith VP, FoV, and horizon line for camera calibration from an input image and a set of line segments. Moreover, two auxiliary outputs (vertical and horizontal convergence line scores) guide the network in learning scene geometry for calibration.

consists of real-world images with labeled horizon lines. SA-MobileNet [150] proposed an image tilt detection and correction with self-attention MobileNet [187] for smartphones. A spatial self-attention module was devised to learn long-range dependencies and global context within the input images. To address the difficulty of the regression task, they trained the network to estimate multiple angles within a narrow interval of the ground truth tilt, penalizing only those values that locate outside this narrow range.

3.3.2 Composite Parameters

Calibrating the composite parameters aims to estimate the intrinsic parameters and extrinsic parameters simultaneously. By jointly estimating composite parameters and training using data from a large-scale panorama dataset [78], Hold-Geoffroy et al. [29] largely outperformed previous independent calibration tasks. Moreover, Hold-Geoffroy et al. [29] performed human perception research in which the participants were asked to evaluate the realism of 3D objects composited with and without accurate calibration. This data was further designed to a new perceptual measure for the calibration errors. In terms of the feature category, Lee et al. [109] and CTRL-C [139] considered both semantic features and geometric cues for camera calibration. They showed how taking use of geometric features, is capable of facilitating the network to comprehend the underlying perspective structure of an image. The pipeline of CTRL-C is illustrated in Figure 3. In recent literature, more applications are jointly studied with camera calibration, for example, single view metrology [51], 3D human pose and shape estimation [48], depth estimation [60], [126], object pose estimation [176], and image reflection removal [128], etc.

Considering the heterogeneousness and visual implicitness of different camera parameters, CPL [163] estimated the parameters using a novel camera projection loss, exploiting the camera model neural network to reconstruct the 3D point cloud. The proposed loss addressed the training imbalance problem by representing different errors of camera parameters in terms of a unified metric.

3.4 Discussion

3.4.1 Technique Summary

The above methods target automatic calibration without manual intervention and scene assumption. Early literature [24], [25] separately studied the intrinsic calibration

or extrinsic calibration. Driven by large-scale datasets and powerful networks, subsequent works [29], [39], [65], [139] considered a comprehensive camera calibration, inferring various parameters and geometric representations. To relieve the difficulty of learning the camera parameters, some works [89], [137], [151], [170] proposed to learn an intermediate representation. In recent literature, more applications are jointly studied with camera calibration [48], [51], [60], [126], [128]. This suggests solving the downstream vision tasks, especially in 3D tasks may require prior knowledge of the image formation model. Moreover, some geometric priors [41] can alleviate the data-starved requirement of deep learning, showing the potential to bridge the gap between the calibration target and semantic features.

It is interesting to find that increasing more extrinsic calibration methods [115], [167], [174] revisited and restored the traditional feature point-based solutions. The standard extrinsics that describe the camera motion contain limited degrees of freedom, and thus some local features can well represent the spatial correspondence. Besides, the network designed for point learning significantly improves the efficiency of calibration models, such as PointNet [188] and PointCNN [189]. Such a pipeline also enables clear interpretability of learning-based camera calibration, which promotes understanding of how the network calibrates and magnifies the influences of intermediate modules.

NeRF, recognized for its groundbreaking capability in synthesizing novel views from 2D images, has seen various advancements recently. Progress in this arena includes the incorporation of additional trainable components, prior constraints, enhanced network designs, and novel training strategies. Specifically, NeRF— [190] marked a pivotal moment, demonstrating the simultaneous optimization of camera parameters and poses during training. It introduced a trainable pinhole camera model, followed by SCNeRF [191] which proposed a more complex camera model, comprising of a pinhole design, radial distortion, and a pixel-specific noise model. While increased trainable parameters offer enhanced representational capabilities, they also complicate training. To counteract this, researchers have utilized prior geometric knowledge as constraints to stabilize optimization, integrating depth estimation [192], [193], multi-view correspondence [191], [193], [194], and GAN-based constraints [195]. Additionally, refinements to NeRF’s architecture have emerged, leveraging Gaussian [196] or sinusoidal activations [197]. Concurrently, training strategies such as coarse-to-fine pipelines [192], [198] and advanced sampling techniques [197], [198], [199] have been proposed, enhancing both reconstruction quality and parameter precision.

Popular frameworks like InstantNGP [200] and NeRF-Studio [201] offer features to fine-tune camera parameters. Typically, NeRF methods use outputs from tools like COLMAP or Polycam, with rendering quality tied closely to initial data quality. Minor errors in intrinsics or poses can lead to noticeable rendering artifacts. Current approaches [200], [201] integrate intrinsic parameters, distortion coefficients, and camera poses into NeRF’s training optimization, enhancing both geometric structure learning and parameter optimization. This boosts 3D geometry robustness and refines camera models. However, fine-tuning parameters may complicate optimization and can destabilize

training. It may also not work well if there are not enough views to reliably adjust the intrinsics without overfitting.

3.4.2 Future Effort

(1) Explore more model priors. Most learning-based calibration methods study the parametric camera models but their generalization abilities are limited. In contrast, non-parametric models directly model the relationship between the 3D imaging ray and its resulting pixel in the image, encoding valuable priors in the learned semantic features to reason the camera parameters. Recent works [202], [203], [204] also incorporate the perspective of modeling pixel-wise information for camera calibration, making minimal assumptions on the camera model and showing more interpretable and in line with how humans perceive.

(2) Decouple different stages in an end-to-end calibration learning model. Most learning-based camera calibration methods include a feature extraction stage and an objective estimation stage. However, how the networks learn the features related to calibration is ambiguous. Therefore, decoupling the learning process by different traditional calibration stages can guide the way of feature extraction. It would be meaningful to extend the idea in extrinsic calibration [115], [167], [174] to more general calibration problems.

(3) Transfer the measurement space from the parameter error to the geometric difference. When it comes to jointly calibrating various camera parameters, the training process will suffer from an imbalance loss optimization problem. The main reason is different camera parameters correspond to different sample distributions. The simple normalization strategy cannot unify their error spaces. Therefore, we can formulate a straightforward measurement space in terms of the geometric properties of different camera parameters.

(4) Despite the strides achieved by recent NeRF methods, training NeRF without precise camera parameters remains a challenge, especially in scenarios with sparse views, pronounced movements, low-texture regions, and suboptimal initial values. Contemporary NeRF-based methodologies do optimize camera parameters, yielding notable results. However, they exhibit significant computational demands and lack the generalization seen in current deep-learning calibration techniques. We contend that in the present NeRF-based methods, camera parameters often play a secondary role. Thus, crafting effective calibration algorithms that capitalize on NeRF remains an arduous but promising endeavor.

4 DISTORTION MODEL

In the learning-based camera calibration, calibrating the radial distortion and roll shutter distortion gains increasing attention due to their widely used applications for the wide-angle lens and CMOS sensor. In this part, we mainly review the calibration/rectification of these two distortions.

4.1 Radial Distortion

The literature on learning-based radial distortion calibration can be classified into two main categories: regression-based solutions and reconstruction-based solutions.

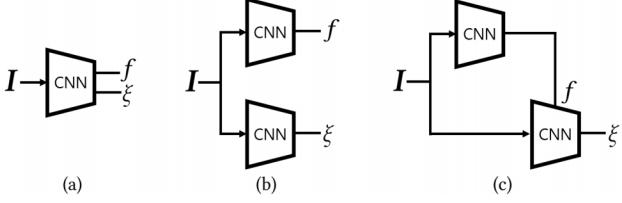


Fig. 4. Three common learning solutions of the regression-based wide-angle camera calibration: (a) SingleNet, (b) DualNet, (c) SeqNet, where I is the distortion image and f and ξ denote the focal length and distortion parameters, respectively. The figure is from [40].

4.1.1 Regression-based Solution

Rong et al. [26] and DeepCalib [40] are pioneer works for the learning-based wide-angle camera calibration. They treated the camera calibration as a supervised classification [26] or regression [40] problem, and then the networks with the convolutional layers and fully connected layers were used to learn the distortion features of inputs and predict the camera parameters. In particular, DeepCalib [40] explored three learning solutions for wide-angle camera calibration as illustrated in Figure 4. Their experiments showed the simplest architecture SingleNet achieves the best performance on both accuracy and efficiency. To enhance the distortion perception of networks, the following works investigated introducing more diverse features such as the semantic features [79] and geometry features [101], [123], [125]. Additionally, some works improved the generalization by designing learning strategies such as unsupervised learning [57], self-supervised learning [59], and reinforcement learning [46]. By randomly chosen coefficients throughout each mini-batch of the training process, RDC-Net [120] was able to dynamically generate distortion images on-the-fly. It enhanced the rectification performance and prevents the learning model from overfitting. Instead of contributing to the techniques of deep learning, other works leaned to explore the vision prior to interpretable calibration. For example, having observed the radial distortion image owns the center symmetry characteristics, in which the texture far from the image center has stronger distortion, Shi et al. [81] and PSE-GAN [119] developed a position-aware weight layer (fixed [81] and learnable [119]) of this property and enabled the network to explicitly perceive the distortion. Lopez et al. [88] proposed a novel parameterization for radial distortion that is better suited for networks than directly learning the distortion parameters. Furthermore, OrdinalDistortion [142] presented a learning-friendly representation, *i.e.*, ordinal distortion. Compared to the implicit and heterogeneous camera parameters, such a representation can facilitate the distortion perception of the neural network due to its clear relation to the image features.

4.1.2 Reconstruction-based Solution

Inspired by the conditional image-to-image translation and dense visual perception, the reconstruction-based solution starts to evolve from the conventional regression-based paradigm. DR-GAN [34] is the first reconstruction-based solution for calibrating the radial distortion, which directly models the pixel-wise mapping between the distorted image and rectified image. It achieved the camera parameter-free training and one-stage rectification. Thanks to the liberation

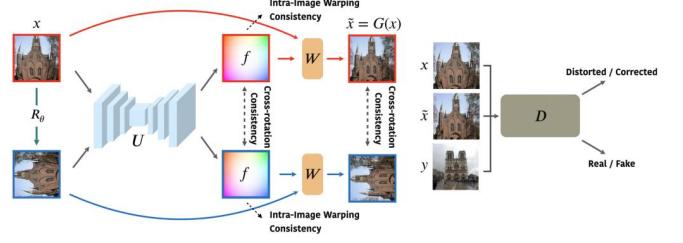


Fig. 5. Architecture of FE-GAN. The figure is from [121]. It consists of two components: a generator $G = (U, W)$ that rectifies the distortion image x , and a discriminator $D = (D_{adv}, D_{cls})$. The module U in G predicts the distortion flow $f = U(x)$, while W rectifies the distortion image using f .

of the assumption of camera models, the reconstruction-based solution showed the potential to calibrate various types of cameras in one learning network. For example, DDM [35] unified different camera models into a domain by presenting the distortion distribution map, which explicitly describes the distortion level of each pixel in a distorted image. Then, the network learned to reconstruct the rectified image using this geometric prior map. To make the mapping function interpretable, the subsequent works [37], [46], [47], [50], [96], [121], [143], [145] developed the displacement field between the distorted image and rectified image. Such a manner is able to eliminate the generated artifacts in the pixel-level reconstruction. In particular, FE-GAN [121] integrated the geometry prior like Shi et al. [81] and PSE-GAN [119] into their reconstruction-based solution and presented a self-supervised strategy to learn the distortion flow for wide-angle camera calibration in Figure 5. Most reconstruction-based solutions exploit a U-Net-like architecture to learn pixel-level mapping. However, the distortion feature can be transferred from encoder to decoder by the skip-connection operation, leading to a blurring appearance and incomplete correction in reconstruction results. To address this issue, Li et al. [117] abandoned the skip-connection in their rectification network. To keep the feature fusion and restrain the geometric difference simultaneously, PCN [145] designed a correction layer in skip-connection and applied the appearance flows to revise the convolved features in different encoder layers. Having noticed that the previous sampling strategy of the convolution kernel neglected the radial symmetry of distortion, PolarRecNet [143] transformed the distorted image from the Cartesian coordinates domain into the polar coordinates domain.

4.2 Roll Shutter Distortion

The existing deep learning calibration works on roll shutter (RS) distortion can be classified into two categories: single-frame-based [27], [42], [100] and multi-frame-based [49], [153], [158], [159], [166]. The single-frame-based solution studies the case of a single roll shutter image as input and directly learns to correct the distortion using neural networks. The ideal corrected result can be regarded as the global shutter (GS) image. It is an ill-posed problem and requires some additional prior assumptions to be defined. On the contrary, the multi-frame-based solution considers the consecutive frames (two or more) of a video taken by a roll shutter camera, in which the strong temporal correlation can be investigated for more reasonable correction.

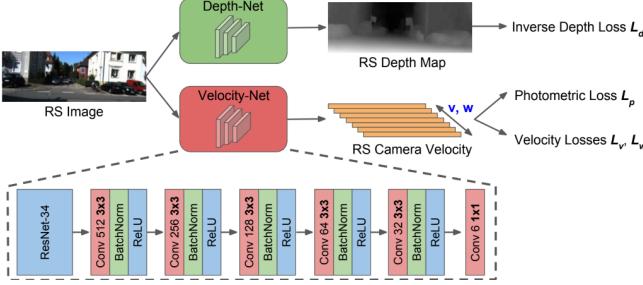


Fig. 6. Architecture of RSC-Net. The figure is from [100]. It consists of two sub-networks, namely DepthNet and Velocity-Net, for learning an RS depth map and RS camera motion from an input image, respectively. Among them, a 6-DOF camera velocity is regressed, including a 3D translational velocity vector v and 3D angular velocity vector w .

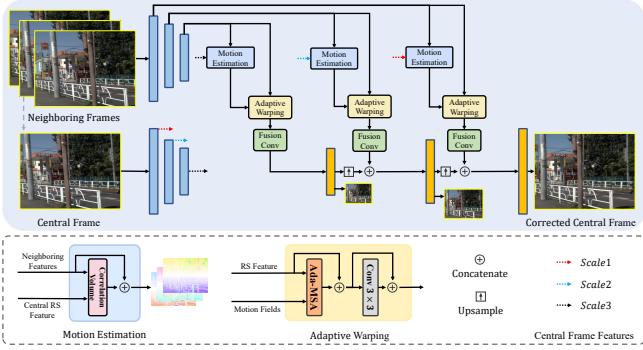


Fig. 7. Architecture of AW-RSC. The figure is from [166]. To address current imprecise motion estimation, it attempts to predict multiple displacement fields instead of only one. Additionally, AW-RSC suggests an adaptive warping module that uses the bundle of fields to guide the adaptive warping of the RS features into the GS one.

4.2.1 Single-frame-based Solution

URS-CNN [27] is the first learning work for calibrating the rolling shutter camera. In this work, a neural network with long kernel characteristics was used to understand how the scene structure and row-wise camera motion interact. To specifically address the nature of the RS effect produced by the row-wise exposure, the row-kernel and column-kernel convolutions were leveraged to extract attributes along horizontal and vertical axes. RSC-Net [100] improved URS-CNN [27] from 2 degrees of freedom (DoF) to 6-DoF and presents a structure-and-motion-aware RS correction model, where the camera scanline velocity and depth were estimated. Compared to URS-CNN [27], RSC-Net [100] further reasoned about the concealed motion between the scanlines as well as the scene structure as shown in Figure 6. To bridge the spatiotemporal connection between RS and GS, EvUnroll [42] exploited the neuromorphic events to correct the RS effect. Event cameras can overcome a number of drawbacks of conventional frame-based activities for dynamic situations with quick motion due to their high temporal resolution property with microsecond-level sensitivity.

4.2.2 Multi-frame-based Solution

Most multi-frame-based solutions are based on the reconstruction paradigm, they mainly devote to contributing how to represent the dense displacement field between RS and global GS images and accurately warp the RS domain to the GS domain. For the first time, DeepUnrollNet [49] proposed

an end-to-end network for two consecutive rolling shutter images using a differentiable forward warping module. In this method, a motion estimation network is used to estimate the dense displacement field from a rolling shutter image to its matching global shutter image. The second contribution of DeepUnrollNet [49] is to construct two novel datasets: the Fastec-RS dataset and the Carla-RS dataset. Furthermore, JCD [153] jointly considered the rolling shutter correction and deblurring (RSCD) techniques, which largely exist in the medium and long exposure cases of rolling shutter cameras. It applied bi-directional warping streams to compensate for the displacement while keeping the non-warped deblurring stream to restore details. The authors also contributed a real-world dataset using a well-designed beam-splitter acquisition system, BS-RSCD, which includes both ego-motion and object motion in dynamic scenes. SUNet [159] extended DeepUnrollNet [49] from the middle time of the second frame ($\frac{3\tau}{2}$) into the intermediate time of two frames (τ). By using PWC-Net [205], SUNet [159] estimated the symmetric undistortion fields and reconstructed the potential GS frames by a time-centered GS image decoder network. To effectively reduce the misalignment between the contexts warped from two consecutive RS images, the context-aware undistortion flow estimator and the symmetric consistency enforcement were designed. To achieve a higher frame rate, Fan et al. [158] generated a GS video from two consecutive RS images based on the scanline-dependent nature of the RS camera. In particular, they first analyzed the inherent connection between bidirectional RS undistortion flow and optical flow, demonstrating the RS undistortion flow map has a more pronounced scanline dependency than the isotropically smooth optical flow map. Then, they developed the bidirectional undistortion flows to describe the pixel-wise RS-aware displacement, and further devised a computation technique for the mutual conversion between different RS undistortion flows corresponding to various scanlines. To eliminate the inaccurate displacement field estimation and error-prone warping problems in previous methods, AW-RSC [166] proposed to predict multiple fields and adaptively warped the learned RS features into global shutter counterparts. Using a coarse-to-fine approach, these warped features were combined and generated to precise global shutter frames as shown in Figure 7. Compared to previous works [49], [153], [158], [159], the warping operation consisting of adaptive multi-head attention and a convolutional block in AW-RSC [166] is learnable and effective. In addition, AW-RSC [166] contributed a real-world rolling shutter correction dataset: BS-RSC, where the RS videos with corresponding GS ground truth are captured simultaneously with a beam-splitter-based acquisition system.

4.3 Discussion

4.3.1 Technique Summary

The deep learning works on wide-angle camera and roll shutter calibration share a similar technique pipeline. Along this research trend, most early literature begins with the regression-based solution [26], [27], [40]. The subsequent works innovated the traditional calibration with a reconstruction perspective [34], [35], [49], [121], which directly learns the displacement field to rectify the uncalibrated

input. For higher accuracy of calibration, a more intuitive displacement field, and more effective warping strategy have been developed [145], [153], [158], [166]. To fit the distribution of different distortions, some works designed different shapes of the convolutional kernel [27] or transformed the convolved coordinates [143].

Existing works devoted themselves to designing more powerful networks and introducing more diverse features to facilitate calibration performance. Increasingly more methods focused on the geometry priors of the distortion [81], [119], [121]. These priors can be directly weighted into the convolutional layers or used to supervise network training, promoting the learning model to converge faster.

4.3.2 Future Effort

(1) The development of wide-angle camera calibration and roll shutter camera calibration can promote each other. For instance, the well-studied multi-frame-based solution in roll shutter calibration is able to inspire wide-angle calibration. The same object located at different sequences could provide useful priors regarding to radial distortion. Additionally, the elaborate studies of the displacement field and warping layer [153], [158], [166] have the potential to motivate the development of wide-angle camera calibration and other fields. Furthermore, the investigation of geometric priors in wide-angle calibration could also improve the interpretability of the network in roll shutter calibration.

(2) Most methods synthesize their training dataset based on random samples from all camera parameters. However, for the images captured by real lenses, the distribution of camera parameters probably locates at a potential manifold [88]. Learning on a label-redundant calibration dataset makes the training process inefficient. Thus, exploring a practical sampling strategy for the synthesized dataset could be a meaningful task in the future direction.

(3) To overcome the ill-posed problem of single-frame calibration, introducing other high-precision sensors can compensate for the current calibration performance, such as event cameras [42]. With the rapid development of vision sensors, joint calibration using multiple sensors is valuable. Consequently, more cross-modal and multi-modal fusion techniques will be investigated along this research way.

5 CROSS-VIEW MODEL

The existing deep calibration methods can estimate the specific camera parameters from a single camera. In fact, there can be more complicated parameter representations in multi-camera circumstances. For example, in the multi-view model, the fundamental matrix and essential matrix describe the epipolar geometry and they are intricately tangled with intrinsics and extrinsics. The homography depicts the pixel-level correspondences between different views. In addition to intrinsics and extrinsics, it is also intertwined with depth. Among these complex parameter representations, homography is the most widely leveraged in practical applications and its related learning-based methods are the most investigated. To this end, we mainly focus on the review of deep homography estimation solutions for the cross-view model and they can be divided into three categories: direct, cascaded, and iterative solution.

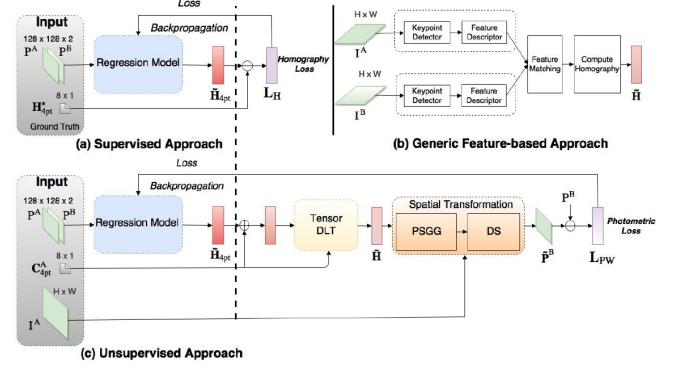


Fig. 8. Architectures of DHN [28] and UDHN [52]. The figure is from [52]. The supervised approach [28] learns to regress a 4 point parameterization of homography \hat{H}_{4pt} using \mathcal{L}_2 loss. The unsupervised approach [52] outputs \hat{H}_{4pt} that minimizes the \mathcal{L}_1 pixel-wise photometric loss of paired inputs (DLT: direct linear transform; PSGG: parameterized sampling grid generator; DS: differentiable sampling).

5.1 Direct Solution

We review the direct deep homography solutions from the perspective of different parameterizations, including the classical 4-pt parameterization and other parameterizations.

5.1.1 4-pt Parameterization

Deep homography estimation is first proposed in DHN [28], where a VGG-style network is adopted to predict the 4-pt parameterization H_{4pt} . To train and evaluate the network, a synthetic dataset named Warped MS-COCO is created to provide ground truth 4-pt parameterization \hat{H}_{4pt} . The pipeline is illustrated in Fig. 8(a), and the objective function is formulated as L_H :

$$L_H = \frac{1}{2} \| H_{4pt} - \hat{H}_{4pt} \|_2^2. \quad (2)$$

Then the 4-pt parameterization can be solved as a 3×3 homography matrix using normalized DLT [206]. However, DHN is limited to synthetic datasets where the ground truth can be generated for free or requires costly labeling of real-world datasets. Subsequently, the first unsupervised solution named UDHN [52] is proposed to address this problem. As shown in Fig. 8(c), it used the same network architecture as DHN and defined an unsupervised loss function by minimizing the average photometric error motivated by traditional methods [207]:

$$L_{PW} = \| \mathcal{P}(I_A(x)) - \mathcal{P}(I_B(\mathcal{W}(x; p))) \|_1, \quad (3)$$

where $\mathcal{W}(\cdot; \cdot)$ and $\mathcal{P}(\cdot)$ denote the operations of warping via homography parameters p and extracting an image patch, respectively. I_A and I_B are the original images with overlapping regions. The input of UDHN is a pair of image patches, but it warps the original images when calculating the loss. In this manner, it avoids the adverse effects of invalid pixels after warping and lifts the magnitude of pixel supervision. To gain accuracy and speed with a tiny model, Chen et al. proposed ShuffleHomoNet [148], which integrates ShuffleNet compressed units [208] and location-aware pooling [84] into a lightweight model. To further handle large displacement, a multi-scale weight-sharing version is exploited by extracting multi-scale feature representations

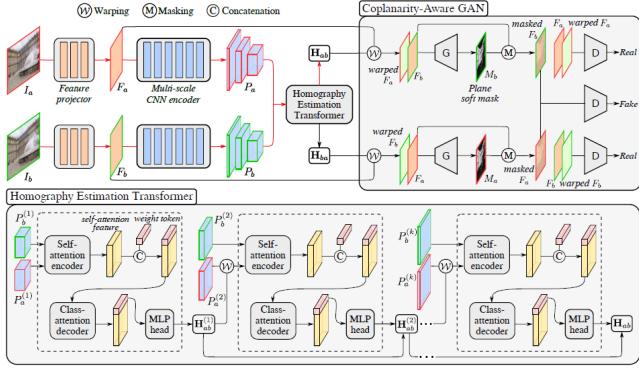


Fig. 9. Architecture of HomoGAN. The figure is from [55]. In particular, the homography estimation transformer with cascaded encoder-decoder blocks takes a feature pyramid of each image as inputs, and predicts the homography from coarse to fine. Coplanarity-aware GAN imposes coplanarity constraints on the model by predicting soft masks of the dominant plane.

and adaptively fusing multi-scale predictions. However, the homography cannot perfectly align images with parallax caused by non-planar structures with non-overlapping camera centers. To deal with parallax, CA-UDHN [53] designs learnable attention masks to overlook the parallax regions, contributing to better background plane alignment. Besides, the 4-pt homography can be extended to meshflow [56] to realize non-planar accurate alignment.

5.1.2 Other Parameterizations

In addition to 4-pt parameterization, the homography can be parameterized as other formulations. To better utilize homography invertibility, Wang et al. proposed SSR-Net [58]. They established the invertibility constraint through a conventional matrix representation in a cyclic manner. Zeng et al. [85] argued that the 4-point parameterization regressed by a fully-connected layer can harm the spatial order of the corners and be susceptible to perturbations, since four points are the minimum requirement to solve the homography. To address these issues, they formulated the parameterization as a perspective field (PF) that models pixel-to-pixel bijection and designed a PFNet. This extends the displacements of the four vertices to as many dense pixel points as possible. The homography can then be solved using RANSAC [209] with outlier filtering, enabling robust estimation by utilizing dense correspondences. Nevertheless, dense correspondences lead to a significant increase in the computational complexity of RANSAC. Furthermore, Ye et al. [54] proposed an 8-DOF flow representation without extra post-processing, which has a size of $H \times W \times 2$ in an 8D subspace constrained by the homography. To represent arbitrary homography flows in this subspace, 8 flow bases are defined, and the proposed BasesHomo is to predict the coefficients for the flow bases. To obtain desirable bases, BasesHomo first generates 8 homography flows by modifying every single entry of an identity homography matrix except for the last entry. Then, these flows are normalized by their largest flow magnitude followed by a QR decomposition, enforcing all the bases normalized and orthogonal.

5.2 Cascaded Solution

Direct solutions explore various homography parameterizations with simple network structures, while the cascaded

ones focus on complex designs of network architectures.

In HierarchicalNet [72], Nowruzi et al. hold that the warped images can be regarded as the input of another network. Therefore they stacked the networks sequentially to reduce the error bounds of the estimate. Based on HierarchicalNet, SRHEN [136] introduced the cost volume [205] to the cascaded network, measuring the feature correlation by cosine distance and formulating it as a volume. The stacked networks and cost volume increase the performance, but they cannot handle the dynamic scenes. MHN [113] developed a multi-scale neural network and proposed to learn homography estimation and dynamic content detection simultaneously. Moreover, to tackle the cross-resolution problem, LocalTrans [147] formulated it as a multimodal problem and proposed a local transformer network embedded within a multiscale structure to explicitly learn correspondences between the multimodal inputs. These inputs include images with different resolutions, and LocalTrans achieved superior performance on cross-resolution cases with a resolution gap of up to 10x. All the solutions mentioned above leverage image pyramids to progressively enhance the ability to address large displacements. However, every image pair at each level requires a unique feature extraction network, resulting in the redundancy of feature maps. To alleviate this problem, some researchers [44], [55], [149], [210] replaced image pyramids with feature pyramids. Specifically, they warped the feature maps directly instead of images to avoid excessive feature extraction networks. To address the low-overlap homography estimation problem in real-world images [149], Nie et al. [149] modified the unsupervised constraint (Eq. 3) to adapt to low-overlap scenes:

$$L'_{PW} = \| I_A(x) \cdot \mathbb{1}(\mathcal{W}(x; p)) - I_B(\mathcal{W}(x; p)) \|_1, \quad (4)$$

where $\mathbb{1}$ is an all-one matrix with the same size as I_A or I_B . It solved the low-overlap problem by taking the original images as network input and ablating the corresponding pixels of I_A to the invalid pixels of warped I_B . To solve the non-planar homography estimation problem, DAMG-Homo [44] proposed backward multi-grid deformation with contextual correlation to align parallax images. Compared with traditional cost volume, the proposed contextual correlation helped to reach better accuracy with lower computational complexity. Another way to address the non-planar problem is to focus on the dominant plane. In HomoGAN [55], an unsupervised GAN is proposed to impose a coplanarity constraint on the predicted homography, as shown in Figure 9. To implement this approach, a generator is used to predict masks of aligned regions, while a discriminator is used to determine whether two masked feature maps were produced by a single homography.

5.3 Iterative Solution

Compared with cascaded methods, iterative solutions achieve higher accuracy by iteratively optimizing the last estimation. Lucas-Kanade (LK) algorithm [207] is usually used in image registration to estimate the parameterized warps iteratively, such as affine transformation, optical flow, etc. It aims at the incremental update of warp parameters Δp every iteration by minimizing the sum of squared error between a template image T and an input image I :

$$E(\Delta p) = \| T(x) - I(\mathcal{W}(x; p + \Delta p)) \|_2^2. \quad (5)$$

However, when optimizing Eq. 5 using first-order Taylor expansion, $\partial I(\mathcal{W}(x; p))/\partial p$ should be recomputed every iteration because $I(\mathcal{W}(x; p))$ varies with p . To avoid this problem, the inverse compositional (IC) LK algorithm [211], an equivalence to LK algorithm, can be used to reformulate the optimization goal as follows:

$$E'(\Delta p) = \| T(\mathcal{W}(x; \Delta p)) - I(\mathcal{W}(x; p)) \|_2^2. \quad (6)$$

After linearizing Eq. 6 with first-order Taylor expansion, we compute $\partial T(\mathcal{W}(x; 0))/\partial p$ instead of $\partial I(\mathcal{W}(x; p))/\partial p$, which would not vary every iteration.

To combine the advantages of deep learning with IC-LK iterator, CLKN [71] conducted LK iterative optimization on semantic feature maps extracted by CNNs as follows:

$$E^f(\Delta p) = \| F_T(\mathcal{W}(x; \Delta p)) - F_I(\mathcal{W}(x; p)) \|_2^2, \quad (7)$$

where F_T and F_I are the feature maps of the template and input images. Then, they enforced the network to run a single iteration with a hinge loss, while the network runs multiple iterations until the stopping condition is met in the testing stage. Besides, CLKN stacked three similar LK networks to further boost the performance by treating the output of the last LK network as the initial warp parameters of the next LK network. From Eq. 7, the IC-LK algorithm heavily relied on feature maps, which tend to fail in multimodal images. Instead, DLKFM [146] constructed a single-channel feature map by using the eigenvalues of the local covariance matrix on the output tensor. To learn DLKFM, it designed two special constraint terms to align multimodal feature maps and contribute to convergence.

However, LK-based algorithms can fail if the Jacobian matrix is rank-deficient [212]. Additionally, the IC-LK iterator is untrainable, which means this drawback is theoretically unavoidable. To address this issue, a completely trainable iterative homography network (IHN) [165] was proposed. Inspired by RAFT [213], IHN updates the cost volume to refine the estimated homography using the same estimator repeatedly every iteration. Furthermore, IHN can handle dynamic scenes by producing an inlier mask in the estimator without requiring extra supervision.

5.4 Discussion

5.4.1 Technique Summary

The above works are devoted to exploring different homography parameterizations such as 4-pt parameterization [28], perspective field [85], and motion bases representation [54], which contributes to better convergence and performance. Other works tend to design various network architectures. In particular, cascaded and iterative solutions are proposed to refine the performance progressively, which can be further combined together to reach higher accuracy. To make the methods more practical, various challenging problems are preliminarily addressed, such as cross resolutions [147], multiple modalities [146], [165], dynamic objects [113], [165], and non-planar scenes [44], [53], [55], etc.

5.4.2 Challenge and Future Effort

We summarize the existing challenges as follows:

(1) Many homography estimation solutions are designed for fixed resolutions, while real-world applications often

involve much more flexible resolutions. When pre-trained models are applied to images with different resolutions, performance can dramatically drop due to the need for input resizing to satisfy the regulated resolution.

(2) Unlike optical flow estimation, which assumes small motions between images, homography estimation often deals with images that have significantly low-overlap rates. In such cases, existing methods may exhibit inferior performance due to limited receptive fields.

(3) Existing methods address the parallax or dynamic objects by learning to reject outliers in the feature extractor [53], cost volume [214], or estimator [165]. However, it is still unclear which stage is more appropriate for outlier rejection.

Based on the challenges we have discussed, some potential research directions for future efforts can be identified:

(1) To overcome the first challenge, we can design various strategies to enhance resolution robustness, such as resolution-related data augmentation, and continual learning on multiple datasets with different resolutions. Besides, we can also formulate a resolution-free parameterization form. The perspective field [85] is a typical case, which represents the homography as dense correspondences with the same resolution as input images. But it requires RANSAC as the post-processing approach, introducing extra computational complexity, especially in the case of extensive correspondences. Therefore, a resolution-free and efficient parameterization form should be explored.

(2) To enhance the performance in low-overlap rate, the main insight is to increase the receptive fields of a network. To this end, the cross-attention module of the transformer explicitly leverages the long-range correlation to eliminate short-range inductive bias [215]. On the other hand, we can exploit beneficial varieties of cost volume to integrate feature correlation [44], [165].

(3) As there is no interaction between different image features in the feature extractor, it is reasonable to assume that outlier rejection should occur after feature extraction. It is not possible to identify outliers within a single image as the depth alone cannot be used as an outlier cue. For example, images captured by purely rotated cameras do not contain parallax outliers. Additionally, it seems intuitive to learn the capability of outlier rejection by combining global and local correlation, similar to the insight of RANSAC.

6 CROSS-SENSOR MODEL

Multi-sensor calibration estimates intrinsic and extrinsic parameters of multiple sensors like cameras, LiDARs, and IMUs. This ensures that data from different sensors are synchronized and registered in a common coordinate system, allowing them to be fused together for a more accurate representation of the environment. Accurate multi-sensor calibration is crucial for applications like autonomous driving and robotics, where reliable sensor fusion is necessary for safe and efficient operation.

In this part, we mainly review the literature on learning-based camera-LiDAR calibration, *i.e.*, predicting the 6-DoF rigid body transformation between a camera and a 3D LiDAR, without requiring any presence of specific features or landmarks in the implementation. Like calibration works on other types of cameras/systems, this research

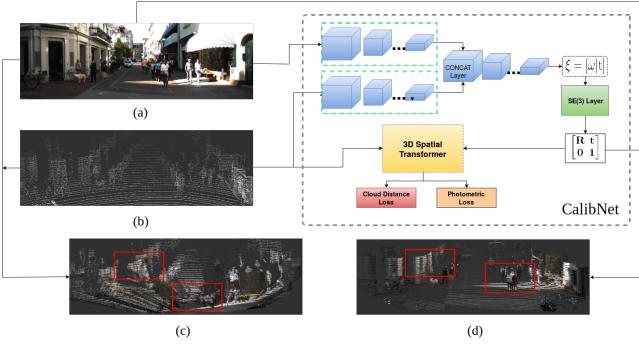


Fig. 10. Network architecture of CalibNet. The figure is from [86]. It takes an RGB image from a calibrated camera and a raw LiDAR point cloud as inputs, and regresses a 6-DoF transformation by an $SE(3)$ layer.

field can also be classified into regression-based solutions and flow/reconstruction-based solutions. But we are prone to follow the special *matching* principle in camera-LiDAR calibration and divide the existing learning-based literature into three categories: pixel-level solution, semantics-level solution, and object/keypoint-level solution.

6.1 Pixel-level Solution

The first deep learning technique in camera-LiDAR calibration, RegNet [30], used CNNs to combine feature extraction, feature matching, and global regression to infer the 6-DoF extrinsic parameters. It processed the RGB and LiDAR depth map separately and branched two parallel data network streams. Then, a specific correlation layer was proposed to convolve the stacked LiDAR and RGB features as a joint representation. After this feature matching, the global information fusion and parameter regression were achieved by two fully connected layers with a Euclidean loss function. Motivated by this work, the subsequent works made a further step into more accurate camera-LiDAR calibration in terms of the geometric constraint [86], [131], temporal correlation [131], loss design [130], feature extraction [182], feature matching [135], [155], feature fusion [182], and calibration representation [156], [179].

For example, as shown in Figure 10, CalibNet [86] designed a network to predict calibration parameters that maximize the geometric and photometric consistency of images and point clouds, solving the underlying physical problem by 3D Spatial Transformers [216]. To refine the calibration model, CalibRCNN [131] presented a synthetic view and an epipolar geometry constraint to measure the photometric and geometric inaccuracies between consecutive frames, of which the temporal information learned by the LSTM network has been investigated in the learning-based camera-LiDAR calibration for the first time. Since the output space of the LiDAR-camera calibration is on the 3D Special Euclidean Group ($SE(3)$) rather than the normal Euclidean space, RGGNet [130] considered Riemannian geometry constraints in the loss function, namely, used a $SE(3)$ geodesic distance equipped with left-invariant Riemannian metrics to optimize the calibration network. LCCNet [155] exploited the cost volume layer to learn the correlation between the image and the depth transformed by the point cloud. Because the depth map ignores the 3D geometric structure of the point cloud, FusionNet [182]

leveraged PointNet++ [217] to directly learn the features from the 3D point cloud. Subsequently, a feature fusion with Ball Query [217] and attention strategy was proposed to effectively fuse the features of images and point clouds.

CFNet [156] first proposed the calibration flow for camera-LiDAR calibration, which represents the deviation between the positions of initial projected 2D points and ground truth. Compared to directly predicting extrinsic parameters, learning the calibration flow helped the network to understand the underlying geometric constraint. To build precise 2D-3D correspondences, CFNet [156] corrected the originally projected points using the estimated calibration flow. Then the efficient Perspective-n-Point (EPnP) algorithm was applied to calculate the final extrinsic parameters by RANSAC. Because RANSAC is nondifferentiable, DXQ-Net [179] further presented a probabilistic model for LiDAR-camera calibration flow, which estimates the uncertainty to measure the quality of LiDAR-camera data association. Then, the differentiable pose estimation module was designed for solving extrinsic parameters, back-propagating the extrinsic error to the flow prediction network.

6.2 Semantics-level Solution

Semantic features can be well learned and represented by deep neural networks. A perfect calibration enables to accurately align the same instance in different sensors. To this end, some works [45], [132], [134], [160] explored to guide the camera-LiDAR calibration with the semantic information. SOIC [134] calibrated and transforms the initialization issue into the semantic centroids' PnP problem using semantic information. Since the 3D semantic centroids of the point cloud and the 2D semantic centroids of the picture cannot match precisely, a matching constraint cost function based on the semantic components was presented. SSI-Calib [132] reformulated the calibration as an optimization problem with a novel calibration quality metric based on semantic features. Then, a non-monotonic subgradient ascent algorithm was proposed to calculate the calibration parameters. Other works utilized the off-the-shelf segmentation networks for point cloud and image, and optimized the calibration parameters by minimizing semantic alignment loss in single-direction [160] and bi-direction [45].

6.3 Object/Keypoint-level Solution

ATOP [181] designed an attention-based object-level matching network, *i.e.*, Cross-Modal Matching Network to explore the overlapped FoV between camera and LiDAR, which facilitated generating the 2D-3D object-level correspondences. 2D and 3D object proposals were detected by YOLOv4 [218] and PointPillar [219]. Then, two cascaded PSO-based algorithms [220] were devised to estimate the calibration extrinsic parameters in the optimization stage. Using the deep declarative network (DDN) [221], RGKCNNet [183] combined the standard neural layer and a PnP solver in the same network, formulating the 2D-3D data association and pose estimation as a bilevel optimization problem. Therefore, both the feature extraction capability of the convolutional layer and the conventional geometric solver can be employed. Microsoft's human keypoint extraction network [222] was applied to detect the 2D-3D matching keypoints. Additionally, RGKCNNet [183] presented a learnable weight layer that

determines the keypoints involved in the solver, enabling the whole pipeline to be trained end-to-end.

6.4 Discussion

6.4.1 Technique Summary

The current method can be briefly classified based on the principle of building 2D and 3D matching, namely, the calibration target. In summary, most pixel-level solutions utilized the end-to-end framework to address this task. While these solutions delivered satisfactory performances on specific datasets, their generalization abilities are limited. Semantics-level and object/keypoint-level methods derived from traditional calibration offered both acceptable performances and generalization abilities. However, they heavily relied on the quality of fore-end feature extraction.

6.4.2 Research Trend

(1) Network architecture is becoming more complex with the use of different structures for feature extraction, matching, and fusion. Current methods employ strategies like multi-scale feature extraction, cross-modal interaction, cost-volume establishment, and confidence-guided fusion.

(2) Directly regressing 6-DoF parameters yields weak generalization ability. To overcome this, intermediate representations like calibration flow have been introduced. Additionally, calibration flow can handle non-rigid transformations that are common in real-world applications.

(3) Traditional methods require specific environments but have well-designed strategies. To balance accuracy and generalization, a combination of geometric solving algorithms and learning methods has been investigated.

6.4.3 Future Effort

(1) Camera-LiDAR calibration methods typically rely on datasets like KITTI, which provide only initial extrinsic parameters. To create a decalibration dataset, researchers add noise transformations to the initial extrinsics, but this approach assumes a fixed position camera-LiDAR system with miscalibration. In real-world applications, the camera-LiDAR relative pose varies, making it challenging to collect large-scale real data with ground truth extrinsics. To address this challenge, generating synthetic camera-LiDAR data using simulation systems could be a valuable solution.

(2) To optimize the combination of networks and traditional solutions, a more compact approach is needed. Current methods mainly use networks as feature extractors, resulting in non-end-to-end pipelines with inadequate feature extraction adjustments for calibration. A deep declarative network (DDN) is a promising framework for making the entire pipeline differentiable. The aggregation of learning and traditional methods can be optimized using DDN.

(3) The most important aspect of camera-LiDAR calibration is 2D-3D matching. To achieve this, the point cloud is commonly transformed into a depth image. However, large deviations in extrinsic simulation can result in detail loss. With the great development of Transformer and cross-modal techniques, we believe leveraging Transformer to directly learn the features of image and point cloud in the same pipeline could facilitate better 2D-3D matching.

7 BENCHMARK

As there is no public and unified benchmark in learning-based camera calibration, we contribute a dataset that can serve as a platform for generalization evaluations. In this dataset, the images and videos are captured by different cameras under diverse scenes, including simulation environments and real-world settings. Additionally, we provide the calibration ground truth, parameter label, and visual clues in this dataset based on different conditions. Figure 11 shows some samples of our collected dataset. Please refer to the evaluation of representative calibration methods on this benchmark in supplementary material.

Standard Model. We collected 300 high-resolution images on the Internet, captured by popular digital cameras such as Canon, Fujifilm, Nikon, Olympus, Sigma, Sony, etc. For each image, we provide the specific focal length of its lens. We have included a diverse range of subjects, including landscapes, portraits, wildlife, architecture, etc. The range of focal length is from 4.5mm to 600mm.

Distortion Model. We created a comprehensive dataset for the distortion camera model, with a focus on wide-angle cameras. The dataset is comprised of three subcategories. The first is a synthetic dataset, which was generated using the widely-used 4th order polynomial model. It contains both circular and rectangular structures, with 1,000 distortion-rectification image pairs. The second subcategory consists of data captured under real-world settings, derived from the raw calibration data for around 40 types of wide-angle cameras. For each calibration data, the intrinsics, extrinsics, and distortion coefficients are available. Finally, we exploit a car equipped with different cameras to capture video sequences. The scenes cover both indoor and outdoor environments, including daytime and nighttime footage.

Cross-View Model. We selected 500 testing samples at random from each of four representative datasets (MS-COCO [28], GoogleEarth [146], GoogleMap [146], CA-Homo [53]) to create a dataset for the cross-view model. It covers a range of scenarios: MS-COCO provides natural synthetic data, GoogleEarth contains aerial synthetic data, and GoogleMap offers multi-modal synthetic data. Parallax is not a factor in these three datasets, while CAHomo provides real-world data with non-planar scenes. To standardize the dataset, we converted all images to a unified format and recorded the matched points between two views. In MS-COCO, GoogleEarth, and GoogleMap, we used four vertices of the images as the matched points. In CAHomo, we identified six matched key points within the same plane.

Cross-Sensor Model. We collected RGB and point cloud data from Apollo [223], DAIR-V2X [224], KITTI [77], KUCL [225], NuScenes [226], and ONCE [227]. Around 300 data pairs with calibration parameters are included in each category. The datasets are captured in different countries to provide enough variety. Each dataset has a different sensor setup, obtaining camera-LiDAR data with varying image resolution, LiDAR scan pattern, and camera-LiDAR relative location. The image resolution ranges from 2448×2048 to 1242×375, while the LiDAR sensors are from Velodyne and Hesai, with 16, 32, 40, 64, and 128 beams. They include not only normal surrounding multi-view images but also small baseline multi-view data. Additionally, we also added

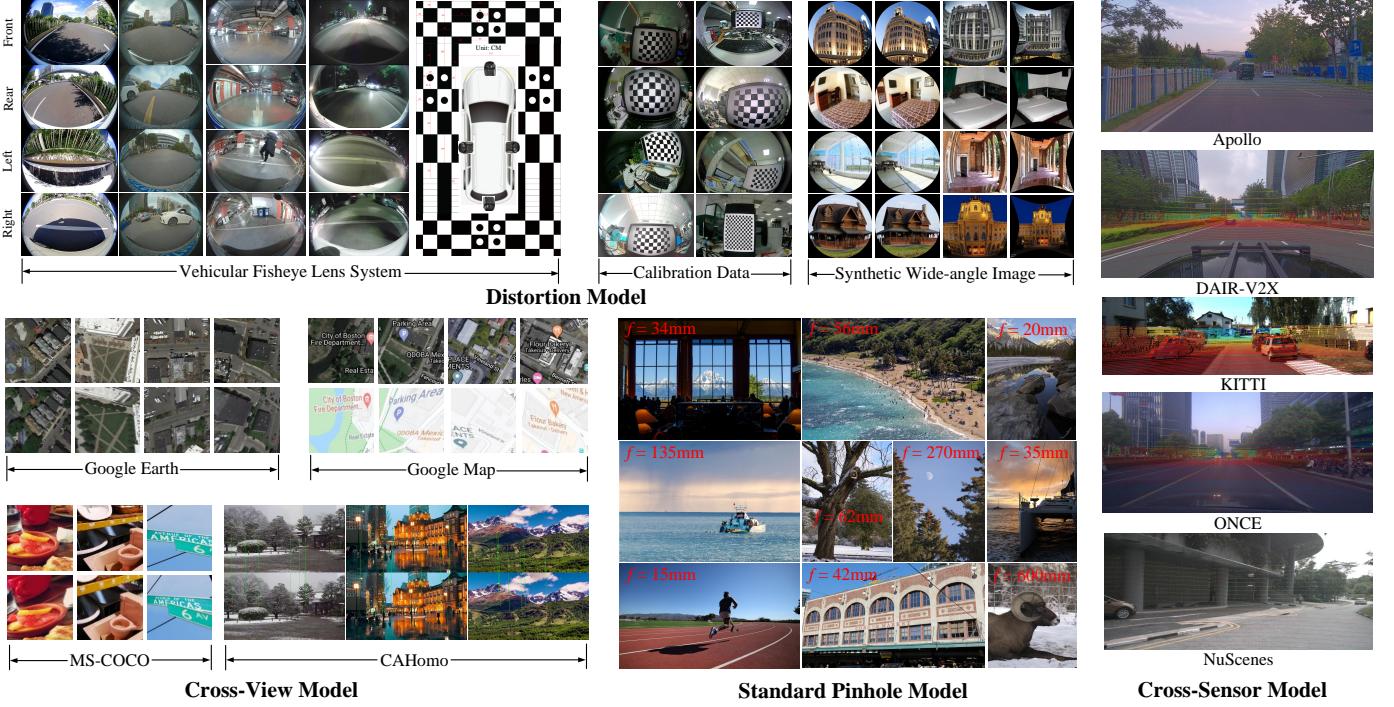


Fig. 11. Overview of our collected benchmark, which covers all models reviewed in this paper. In this dataset, the image and video derive from diverse cameras under different environments. The accurate ground truth and label are provided for each sample.

random disturbance of around 20 degrees rotation and 1.5 meters translation based on classical settings [30] to simulate vibration and collision.

8 FUTURE RESEARCH DIRECTIONS

Camera calibration is a fundamental and challenging research topic. From the above technical reviews and limitation analysis, we can conclude there is still room for improvement with deep learning. From Section 3 to Section 6, specific future efforts are discussed for each model. In this section, we suggest more general future research directions.

8.1 Sequences

Bundle adjustment is a well-established technique central to Multi-View Stereo (MVS) and Simultaneous Localization and Mapping (SLAM) using multi-view constraints. Traditional bundle adjustment focuses on pose estimation, often under the assumption of pre-calibrated cameras, thus sidelining the nuances of camera parameter fine-tuning. While learning-based camera calibration has made significant strides, most methods are tailored for a single image. We have highlighted intrinsics calibration to underscore how sequence constraints bolster prediction accuracy. Notably, there is a burgeoning interest in integrating bundle adjustment into end-to-end deep learning pipelines. By transitioning from conventional keypoint extraction and matching to learning-based methods, recent works [228], [229], [230], [231], [232] propose differentiable bundle adjustment layers to refine pose, depth, and camera parameters together. Consequently, there is immense potential in further harnessing sequence constraints for accurate calibration. Current methods combine front-end matching with a back-end solver, which can be inefficient and unreliable in cases

like fast motion. We suggest separating front-end and back-end refinements, using large models for features and introducing more trainable parameters in optimization.

8.2 Learning Target

Due to the implicit relationship to image features, conventional calibration objectives can be challenging for neural networks to learn. To this end, some works have developed novel learning targets that replace conventional calibration objectives, providing learning-friendly representations for neural networks. Additionally, intermediate geometric representations have been presented to bridge the gap between image features and calibration objectives, such as reflective amplitude coefficient maps [128], rectification flow [37], surface geometry [89], and normal flow [170], etc. Looking ahead to the future development of this community, we believe there is still great potential for designing more explicit and reasonable learning targets for calibration objectives.

8.3 Pre-training

Pre-training on ImageNet [69] has become a widely used strategy in deep learning. However, recent studies [96] have shown that this approach provides less benefit for specific camera calibration tasks, such as wide-angle camera calibration. This is due to two main reasons: the data gap and the task gap. The ImageNet dataset only contains perspective images without distortions, making the initialized weights of neural networks irrelevant to distortion models. Furthermore, He et al. [233] demonstrated that the task of ImageNet pre-training has limited benefits when the final task is more sensitive to localization. As a result, the performance of extrinsics estimation may be impacted by this task gap. Moreover, pre-training beyond a single image and a single modality, to our knowledge, has not been thoroughly

investigated in the related field. We suggest that designing a customized pre-training strategy for learning-based camera calibration is an interesting area of research.

8.4 Implicit Unified Model

Deep learning-based camera calibration methods use traditional parametric camera models, which lack the flexibility to fit complex situations. Non-parametric camera models relate each pixel to its corresponding 3D observation ray, overcoming parametric model limitations. However, they require strict calibration targets and are more complex for undistortion, projection, and unprojection. Deep learning methods show potential for calibration tasks, making non-parametric models worth revisiting and potentially replacing parametric models in the future. Moreover, they allow for implicit and unified calibration, fitting all camera types through pixel-level regression and avoiding explicit feature extraction and geometry solving. Researchers combined the advantages of implicit and unified representation with the Neural Radiance Field (NeRF) for reconstructing 3D structures and synthesizing novel views. Self-calibration NeRF [191] has been proposed for generic cameras with arbitrary non-linear distortions, and end-to-end pipelines have been explored to learn depth and ego-motion without calibration targets. We believe the implicit and unified camera models could be used to optimize learning-based algorithms or integrated into downstream 3D vision tasks.

9 CONCLUSION

In this paper, we present a comprehensive survey of the recent efforts in the area of deep learning-based camera calibration. Our survey covers conventional camera models, classified learning paradigms and learning strategies, detailed reviews of the state-of-the-art approach, a public benchmark, and future research directions. To exhibit the development process and link the connections between existing works, we provide a fine-grained taxonomy that categorizes literature by jointly considering camera models and applications. Moreover, the relationships, strengths, distinctions, and limitations are thoroughly discussed in each category. An open-source repository will keep updating regularly with new works and datasets. We hope that this survey could promote future research in this field.

ACKNOWLEDGMENT

We thank Leidong Qin and Shangrong Yang at Beijing Jiaotong University for the partial dataset collection. We thank Jinlong Fan at the University of Sydney for the insightful discussion.

REFERENCES

- [1] C. B. Duane, "Close-range camera calibration," *Photogramm. Eng.*, vol. 37, no. 8, pp. 855–866, 1971.
- [2] S. J. Maybank and O. D. Faugeras, "A theory of self-calibration of a moving camera," *International Journal of Computer Vision*, vol. 8, no. 2, pp. 123–151, 1992.
- [3] J. Weng, P. Cohen, M. Herniou *et al.*, "Camera calibration with distortion models and accuracy evaluation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, no. 10, pp. 965–980, 1992.
- [4] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 11, pp. 1330–1334, 2000.
- [5] D. C. Brown, "Decentering distortion of lenses," *Photogrammetric Engineering and Remote Sensing*, 1966.
- [6] Z. Zhang, "Flexible camera calibration by viewing a plane from unknown orientations," in *International Conference on Computer Vision*, vol. 1, 1999, pp. 666–673.
- [7] S. Gasparini, P. Sturm, and J. P. Barreto, "Plane-based calibration of central catadioptric cameras," in *International Conference on Computer Vision*, 2009, pp. 1195–1202.
- [8] S. Shah and J. Aggarwal, "A simple calibration procedure for fish-eye (high distortion) lens camera," in *Proceedings of IEEE International Conference on Robotics and Automation*, 1994, pp. 3422–3427.
- [9] J. P. Barreto and H. Araujo, "Geometric properties of central catadioptric line images and their application in calibration," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1327–1333, 2005.
- [10] R. Carroll, M. Agrawal, and A. Agarwala, "Optimizing content-preserving projections for wide-angle images," in *ACM Transactions on Graphics*, vol. 28, no. 3, 2009, p. 43.
- [11] F. Bukhari and M. N. Dailey, "Automatic radial distortion estimation from a single image," *Journal of Mathematical Imaging and Vision*, vol. 45, no. 1, pp. 31–45, 2013.
- [12] M. Aleman-Flores, L. Alvarez, L. Gomez, and D. Santana-Cedrés, "Automatic lens distortion correction using one-parameter division models," *Image Processing On Line*, vol. 4, pp. 327–343, 2014.
- [13] O. D. Faugeras, Q.-T. Luong, and S. J. Maybank, "Camera self-calibration: Theory and experiments," in *European Conference on Computer Vision*, 1992, pp. 321–334.
- [14] C. S. Fraser, "Digital camera self-calibration," *ISPRS Journal of Photogrammetry and Remote sensing*, vol. 52, no. 4, pp. 149–159, 1997.
- [15] R. I. Hartley, "Self-calibration from multiple views with a rotating camera," in *European Conference on Computer Vision*, 1994, pp. 471–478.
- [16] F. Camposeco, T. Sattler, and M. Pollefeys, "Non-parametric structure-based calibration of radially symmetric cameras," in *International Conference on Computer Vision*, 2015, pp. 2192–2200.
- [17] T. Schops, V. Larsson, M. Pollefeys, and T. Sattler, "Why having 10,000 parameters in your camera model is better than twelve," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 2535–2544.
- [18] L. Pan, M. Pollefeys, and V. Larsson, "Camera pose estimation using implicit distortion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12819–12828.
- [19] [Online]. Available: https://docs.opencv.org/4.x/dc/dbb/tutorial_py_calibration.html
- [20] [Online]. Available: <https://www.mathworks.com/help/vision/camera-calibration.html>
- [21] J. Salvi, X. Armangué, and J. Batlle, "A comparative review of camera calibrating methods with accuracy evaluation," *Pattern Recognition*, vol. 35, no. 7, pp. 1617–1635, 2002.
- [22] C. Hughes, M. Glavin, E. Jones, and P. Denny, "Review of geometric distortion compensation in fish-eye cameras," 2008.
- [23] J. Fan, J. Zhang, S. J. Maybank, and D. Tao, "Wide-angle image rectification: a survey," *International Journal of Computer Vision*, vol. 130, no. 3, pp. 747–776, 2022.
- [24] S. Workman, C. Greenwell, M. Zhai, R. Baltenberger, and N. Jacobs, "Deepfocal: A method for direct focal length estimation," in *IEEE International Conference on Image Processing*, 2015, pp. 1369–1373.
- [25] A. Kendall, M. Grimes, and R. Cipolla, "Posenet: A convolutional network for real-time 6-dof camera relocalization," in *International Conference on Computer Vision*, 2015.
- [26] J. Rong, S. Huang, Z. Shang, and X. Ying, "Radial lens distortion correction using convolutional neural networks trained with synthesized images," in *Asian Conference on Computer Vision*, 2016, pp. 35–49.
- [27] V. Rengarajan, Y. Balaji, and A. Rajagopalan, "Unrolling the shutter: CNN to correct motion distortions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2291–2299.
- [28] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Deep image homography estimation," *arXiv preprint arXiv:1606.03798*, 2016.
- [29] Y. Hold-Geoffroy, K. Sunkavalli, J. Eisenmann, M. Fisher, E. Gambarotto, S. Hadap, and J.-F. Lalonde, "A perceptual measure for

- deep single image camera calibration," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.
- [30] N. Schneider, F. Piewak, C. Stiller, and U. Franke, "Regnet: Multimodal sensor registration using deep neural networks," in *IEEE intelligent vehicles symposium*, 2017, pp. 1803–1810.
- [31] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1125–1134.
- [32] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.
- [33] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," *Advances in Neural Information Processing Systems*, vol. 27, 2014.
- [34] K. Liao, C. Lin, Y. Zhao, and M. Gabbouj, "Dr-gan: Automatic radial distortion rectification using conditional gan in real-time," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 3, pp. 725–733, 2020.
- [35] K. Liao, C. Lin, Y. Zhao, and M. Xu, "Model-free distortion rectification framework bridged by distortion distribution map," *IEEE Transactions on Image Processing*, vol. 29, pp. 3707–3718, 2020.
- [36] K. Liao, C. Lin, L. Liao, Y. Zhao, and W. Lin, "Multi-level curriculum for training a distortion-aware barrel distortion rectification model," in *International Conference on Computer Vision*, 2021, pp. 4389–4398.
- [37] X. Li, B. Zhang, P. V. Sander, and J. Liao, "Blind geometric distortion correction on images through deep learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.
- [38] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015, pp. 234–241.
- [39] M. Zhai, S. Workman, and N. Jacobs, "Detecting vanishing points using global image context in a non-manhattan world," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.
- [40] O. Bogdan, V. Eckstein, F. Rameau, and J.-C. Bazin, "Deepcalib: a deep learning approach for automatic intrinsic calibration of wide field-of-view cameras," in *Proceedings of the 15th ACM SIGGRAPH European Conference on Visual Media Production*, 2018.
- [41] Y. Lin, R. Wiersma, S. L. Pintea, K. Hildebrandt, E. Eisemann, and J. C. van Gemert, "Deep vanishing point detection: Geometric priors make dataset variations vanish," *arXiv preprint arXiv:2203.08586*, 2022.
- [42] X. Zhou, P. Duan, Y. Ma, and B. Shi, "Evunroll: Neuromorphic events based rolling shutter image correction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17775–17784.
- [43] Y. Shangrong, L. Chunyu, L. Kang, and Z. Yao, "Fishformer: Annulus slicing-based transformer for fisheye rectification with efficacy domain exploration," *arXiv preprint arXiv:2207.01925*, 2022.
- [44] L. Nie, C. Lin, K. Liao, S. Liu, and Y. Zhao, "Depth-aware multi-grid deep homography estimation with contextual correlation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 7, pp. 4460–4472, 2021.
- [45] K. Akio, Z. Yiyang, Z. Pengwei, Z. Wei, and T. Masayoshi, "Sst-calib: Simultaneous spatial-temporal parameter calibration between lidar and camera," *arXiv preprint arXiv:2207.03704*, 2022.
- [46] Y. Zhao, Z. Huang, T. Li, W. Chen, C. LeGendre, X. Ren, A. Shapiro, and H. Li, "Learning perspective undistortion of portraits," in *International Conference on Computer Vision*, 2019.
- [47] J. Tan, S. Zhao, P. Xiong, J. Liu, H. Fan, and S. Liu, "Practical wide-angle portraits correction with deep structured models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3498–3506.
- [48] M. Kocabas, C.-H. P. Huang, J. Tesch, L. Müller, O. Hilliges, and M. J. Black, "Spec: Seeing people in the wild with an estimated camera," in *International Conference on Computer Vision*, 2021, pp. 11035–11045.
- [49] P. Liu, Z. Cui, V. Larsson, and M. Pollefeys, "Deep shutter unrolling network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5941–5949.
- [50] F. Zhu, S. Zhao, P. Wang, H. Wang, H. Yan, and S. Liu, "Semi-supervised wide-angle portraits correction by multi-scale transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 19689–19698.
- [51] R. Zhu, X. Yang, Y. Hold-Geoffroy, F. Perazzi, J. Eisenmann, K. Sunkavalli, and M. Chandraker, "Single view metrology in the wild," in *European Conference on Computer Vision*, 2020, pp. 316–333.
- [52] T. Nguyen, S. W. Chen, S. S. Shivakumar, C. J. Taylor, and V. Kumar, "Unsupervised deep homography: A fast and robust homography estimation model," *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 2346–2353, 2018.
- [53] J. Zhang, C. Wang, S. Liu, L. Jia, N. Ye, J. Wang, J. Zhou, and J. Sun, "Content-aware unsupervised deep homography estimation," in *European Conference on Computer Vision*, 2020, pp. 653–669.
- [54] N. Ye, C. Wang, H. Fan, and S. Liu, "Motion basis learning for unsupervised deep homography estimation with subspace projection," in *International Conference on Computer Vision*, 2021, pp. 13117–13125.
- [55] M. Hong, Y. Lu, N. Ye, C. Lin, Q. Zhao, and S. Liu, "Unsupervised homography estimation with coplanarity-aware gan," *arXiv preprint arXiv:2205.03821*, 2022.
- [56] S. Liu, N. Ye, C. Wang, J. Zhang, L. Jia, K. Luo, J. Wang, and J. Sun, "Content-aware unsupervised deep homography estimation and its extensions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 3, pp. 2849–2863, 2022.
- [57] S. Yang, C. Lin, K. Liao, Y. Zhao, and M. Liu, "Unsupervised fish-eye image correction through bidirectional loss with geometric prior," *Journal of Visual Communication and Image Representation*, vol. 66, p. 102692, 2020.
- [58] X. Wang, C. Wang, B. Liu, X. Zhou, L. Zhang, J. Zheng, and X. Bai, "Multi-view stereo in the deep learning era: A comprehensive review," *Displays*, vol. 70, p. 102102, 2021.
- [59] J. Fan, J. Zhang, and D. Tao, "Sir: Self-supervised image rectification via seeing the same scene from multiple different lenses," *IEEE Transactions on Image Processing*, 2022.
- [60] J. Fang, I. Vasiljevic, V. Guzilini, R. Ambrus, G. Shakhnarovich, A. Gaidon, and M. R. Walter, "Self-supervised camera self-calibration from video," *arXiv preprint arXiv:2112.03325*, 2021.
- [61] J. Zhao, S. Wei, L. Liao, and Y. Zhao, "Dqn-based gradual fisheye image rectification," *Pattern Recognition Letters*, vol. 152, pp. 129–134, 2021.
- [62] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski et al., "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.
- [63] K. Wilson and N. Snavely, "Robust global translations with 1dsfm," in *European Conference on Computer Vision*, 2014, pp. 61–75.
- [64] [Online]. Available: <https://www.repository.cam.ac.uk/handle/1810/251342?jsessionid=90AB1617B8707CD387CBF67437683F77>
- [65] S. Workman, M. Zhai, and N. Jacobs, "Horizon lines in the wild," *arXiv preprint arXiv:1604.02129*, 2016.
- [66] [Online]. Available: <https://mvrl.cse.wustl.edu/datasets/hlw/>
- [67] P. Denis, J. H. Elder, and F. J. Estrada, "Efficient edge-based methods for estimating manhattan frames in urban imagery," in *European Conference on Computer Vision*, 2008, pp. 197–210.
- [68] O. Barinova, V. Lempitsky, E. Tretiak, and P. Kohli, "Geometric image parsing in man-made environments," in *European Conference on Computer Vision*, 2010, pp. 57–70.
- [69] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.
- [70] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*, 2014, pp. 740–755.
- [71] C.-H. Chang, C.-N. Chou, and E. Y. Chang, "Clkn: Cascaded lucas-kanade networks for image alignment," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.
- [72] F. Erlik Nowruzi, R. Laganiere, and N. Japkowicz, "Homography estimation from image pairs with hierarchical convolutional networks," in *International Conference on Computer Vision Workshops*, 2017.

- [73] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "Sun database: Large-scale scene recognition from abbey to zoo," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010, pp. 3485–3492.
- [74] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1–8.
- [75] H. Shao, T. Svoboda, and L. Van Gool, "Zubud-zurich buildings database for image based recognition," *Computer Vision Lab, Swiss Federal Institute of Technology*, vol. 260, no. 20, p. 6, 2003.
- [76] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "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*, 2008.
- [77] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 3354–3361.
- [78] J. Xiao, K. A. Ehinger, A. Oliva, and A. Torralba, "Recognizing scene viewpoint using panoramic place representation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2695–2702.
- [79] X. Yin, X. Wang, J. Yu, M. Zhang, P. Fua, and D. Tao, "Fisheyerecnet: A multi-context collaborative deep network for fisheye image rectification," in *European Conference on Computer Vision*, 2018.
- [80] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ade20k dataset," in *Proceedings of the IEEE Computer Vision and Pattern Recognition*, 2017, pp. 633–641.
- [81] Y. Shi, D. Zhang, J. Wen, X. Tong, X. Ying, and H. Zha, "Radial lens distortion correction by adding a weight layer with inverted foveal models to convolutional neural networks," in *International Conference on Pattern Recognition*, 2018.
- [82] R. Ranftl and V. Koltun, "Deep fundamental matrix estimation," in *European Conference on Computer Vision*, 2018.
- [83] A. Knapitsch, J. Park, Q.-Y. Zhou, and V. Koltun, "Tanks and temples: Benchmarking large-scale scene reconstruction," *ACM Transactions on Graphics*, vol. 36, no. 4, pp. 1–13, 2017.
- [84] O. Poursaeed, G. Yang, A. Prakash, Q. Fang, H. Jiang, B. Hariharan, and S. Belongie, "Deep fundamental matrix estimation without correspondences," in *European Conference on Computer Vision Workshops*, 2018.
- [85] R. Zeng, S. Denman, S. Sridharan, and C. Fookes, "Rethinking planar homography estimation using perspective fields," in *Asian Conference on Computer Vision*, 2018, pp. 571–586.
- [86] G. Iyer, R. K. Ram, J. K. Murthy, and K. M. Krishna, "Calibnet: Geometrically supervised extrinsic calibration using 3d spatial transformer networks," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2018, pp. 1110–1117.
- [87] C.-K. Chang, J. Zhao, and L. Itti, "Deepvp: Deep learning for vanishing point detection on 1 million street view images," in *IEEE International Conference on Robotics and Automation*, 2018, pp. 4496–4503.
- [88] M. Lopez, R. Mari, P. Gargallo, Y. Kuang, J. Gonzalez-Jimenez, and G. Haro, "Deep single image camera calibration with radial distortion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.
- [89] W. Xian, Z. Li, M. Fisher, J. Eisenmann, E. Shechtman, and N. Snavely, "Uprightnet: Geometry-aware camera orientation estimation from single images," in *International Conference on Computer Vision*, 2019.
- [90] W. Li, S. Saeedi, J. McCormac, R. Clark, D. Tzoumanikas, Q. Ye, Y. Huang, R. Tang, and S. Leutenegger, "Interiornet: Mega-scale multi-sensor photo-realistic indoor scenes dataset," *arXiv preprint arXiv:1809.00716*, 2018.
- [91] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "Scannet: Richly-annotated 3d reconstructions of indoor scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5828–5839.
- [92] B. Zhuang, Q.-H. Tran, G. H. Lee, L. F. Cheong, and M. Chandraker, "Degeneracy in self-calibration revisited and a deep learning solution for uncalibrated slam," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2019, pp. 3766–3773.
- [93] S. Ammar Abbas and A. Zisserman, "A geometric approach to obtain a bird's eye view from an image," in *International Conference on Computer Vision Workshops*, 2019.
- [94] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "Carla: An open urban driving simulator," in *Conference on Robot Learning*, 2017, pp. 1–16.
- [95] O. Barinova, V. Lempitsky, E. Tretiak, and P. Kohli, "Geometric image parsing in man-made environments," in *European Conference on Computer Vision*, 2010, pp. 57–70.
- [96] K. Liao, C. Lin, Y. Zhao, and M. Gabbouj, "Distortion rectification from static to dynamic: A distortion sequence construction perspective," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 11, pp. 3870–3882, 2020.
- [97] R. Jung, A. S. J. Lee, A. Ashtari, and J.-C. Bazin, "Deep360up: A deep learning-based approach for automatic vr image upright adjustment," in *IEEE Conference on Virtual Reality and 3D User Interfaces*, 2019.
- [98] V. Belagiannis, C. Rupprecht, G. Carneiro, and N. Navab, "Robust optimization for deep regression," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2830–2838.
- [99] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 6, pp. 1452–1464, 2018.
- [100] B. Zhuang, Q.-H. Tran, P. Ji, L.-F. Cheong, and M. Chandraker, "Learning structure-and-motion-aware rolling shutter correction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.
- [101] Z. Xue, N. Xue, G.-S. Xia, and W. Shen, "Learning to calibrate straight lines for fisheye image rectification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.
- [102] K. Huang, Y. Wang, Z. Zhou, T. Ding, S. Gao, and Y. Ma, "Learning to parse wireframes in images of man-made environments," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 626–635.
- [103] S. Song, F. Yu, A. Zeng, A. X. Chang, M. Savva, and T. Funkhouser, "Semantic scene completion from a single depth image," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1746–1754.
- [104] L. Yin, X. Sun, T. Worm, and M. Reale, "A high-resolution 3d dynamic facial expression database, 2008," in *IEEE International Conference on Automatic Face and Gesture Recognition, Amsterdam, The Netherlands*, vol. 126.
- [105] Y. Zhou, H. Qi, J. Huang, and Y. Ma, "Neurvps: Neural vanishing point scanning via conic convolution," *Advances in Neural Information Processing Systems*, vol. 32, 2019.
- [106] Y. Zhou, H. Qi, Y. Zhai, Q. Sun, Z. Chen, L.-Y. Wei, and Y. Ma, "Learning to reconstruct 3d manhattan wireframes from a single image," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7698–7707.
- [107] L. Sha, J. Hobbs, P. Felsen, X. Wei, P. Lucey, and S. Ganguly, "End-to-end camera calibration for broadcast videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.
- [108] N. Homayounfar, S. Fidler, and R. Urtasun, "Sports field localization via deep structured models," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5212–5220.
- [109] J. Lee, M. Sung, H. Lee, and J. Kim, "Neural geometric parser for single image camera calibration," in *European Conference on Computer Vision*, 2020, pp. 541–557.
- [110] [Online]. Available: <https://developers.google.com/maps/>
- [111] A. Cramariuc, A. Petrov, R. Suri, M. Mittal, R. Siegwart, and C. Cadena, "Learning camera miscalibration detection," in *IEEE International Conference on Robotics and Automation*, 2020, pp. 4997–5003.
- [112] C. Zhang, F. Rameau, J. Kim, D. M. Argaw, J.-C. Bazin, and I. S. Kweon, "Deepptz: Deep self-calibration for ptz cameras," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020.
- [113] H. Le, F. Liu, S. Zhang, and A. Agarwala, "Deep homography estimation for dynamic scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.
- [114] B. Davidson, M. S. Alvi, and J. F. Henriques, "360° camera alignment via segmentation," in *European Conference on Computer Vision*, 2020, pp. 579–595.
- [115] Y.-Y. Jau, R. Zhu, H. Su, and M. Chandraker, "Deep keypoint-based camera pose estimation with geometric constraints," in

- 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2020, pp. 4950–4957.
- [116] X. Huang, P. Wang, X. Cheng, D. Zhou, Q. Geng, and R. Yang, “The apolloscape open dataset for autonomous driving and its application,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 10, pp. 2702–2719, 2019.
- [117] Y.-H. Li, I.-C. Lo, and H. H. Chen, “Deep face rectification for 360° dual-fisheye cameras,” *IEEE Transactions on Image Processing*, vol. 30, pp. 264–276, 2021.
- [118] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, “Ms-celeb-1m: A dataset and benchmark for large-scale face recognition,” in *European Conference on Computer Vision*, 2016, pp. 87–102.
- [119] Y. Shi, X. Tong, J. Wen, H. Zhao, X. Ying, and H. Zha, “Position-aware and symmetry enhanced gan for radial distortion correction,” in *International Conference on Pattern Recognition*, 2021, pp. 1701–1708.
- [120] H. Zhao, Y. Shi, X. Tong, X. Ying, and H. Zha, “A simple yet effective pipeline for radial distortion correction,” in *IEEE International Conference on Image Processing*, 2020, pp. 878–882.
- [121] C.-H. Chao, P.-L. Hsu, H.-Y. Lee, and Y.-C. F. Wang, “Self-supervised deep learning for fisheye image rectification,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020, pp. 2248–2252.
- [122] F. Yu, A. Seff, Y. Zhang, S. Song, T. Funkhouser, and J. Xiao, “Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop,” *arXiv preprint arXiv:1506.03365*, 2015.
- [123] H. Zhao, X. Ying, Y. Shi, X. Tong, J. Wen, and H. Zha, “Rdcface: Radial distortion correction for face recognition,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.
- [124] F. Wang, L. Chen, C. Li, S. Huang, Y. Chen, C. Qian, and C. C. Loy, “The devil of face recognition is in the noise,” in *European Conference on Computer Vision*, 2018, pp. 765–780.
- [125] Z.-C. Xue, N. Xue, and G.-S. Xia, “Fisheye distortion rectification from deep straight lines,” *arXiv preprint arXiv:2003.11386*, 2020.
- [126] M. Baradad and A. Torralba, “Height and uprightness invariance for 3d prediction from a single view,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.
- [127] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, “Indoor segmentation and support inference from rgbd images,” in *European Conference on Computer Vision*, 2012, pp. 746–760.
- [128] Q. Zheng, J. Chen, Z. Lu, B. Shi, X. Jiang, K.-H. Yap, L.-Y. Duan, and A. C. Kot, “What does plate glass reveal about camera calibration?” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.
- [129] [Online]. Available: <https://figshare.com/articles/dataset/FocaLens/3399169/2>
- [130] K. Yuan, Z. Guo, and Z. J. Wang, “Rggnet: Tolerance aware lidar-camera online calibration with geometric deep learning and generative model,” *IEEE Robotics and Automation Letters*, vol. 5, no. 4, pp. 6956–6963, 2020.
- [131] J. Shi, Z. Zhu, J. Zhang, R. Liu, Z. Wang, S. Chen, and H. Liu, “Calibrccnn: Calibrating camera and lidar by recurrent convolutional neural network and geometric constraints,” in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2020, pp. 10197–10202.
- [132] Y. Zhu, C. Li, and Y. Zhang, “Online camera-lidar calibration with sensor semantic information,” in *IEEE International Conference on Robotics and Automation*, 2020, pp. 4970–4976.
- [133] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, “The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results,” <http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html>.
- [134] W. Wang, S. Nobuhara, R. Nakamura, and K. Sakurada, “Soic: Semantic online initialization and calibration for lidar and camera,” *arXiv preprint arXiv:2003.04260*, 2020.
- [135] S. Wu, A. Hadachi, D. Vivet, and Y. Prabhakar, “Netcalib: A novel approach for lidar-camera auto-calibration based on deep learning,” in *International Conference on Pattern Recognition*, 2021, pp. 6648–6655.
- [136] Y. Li, W. Pei, and Z. He, “Srhen: stepwise-refining homography estimation network via parsing geometric correspondences in deep latent space,” in *Proceedings of the ACM International Conference on Multimedia*, 2020, pp. 3063–3071.
- [137] Y. Gil, S. Elmalem, H. Haim, E. Marom, and R. Giryes, “Online training of stereo self-calibration using monocular depth estimation,” *IEEE Transactions on Computational Imaging*, vol. 7, pp. 812–823, 2021.
- [138] [Online]. Available: <http://www.cs.toronto.edu/~harel/TAUAgent/download.html>
- [139] J. Lee, H. Go, H. Lee, S. Cho, M. Sung, and J. Kim, “Ctrl-c: Camera calibration transformer with line-classification,” in *International Conference on Computer Vision*, 2021, pp. 16228–16237.
- [140] N. Wakai and T. Yamashita, “Deep single fisheye image camera calibration for over 180-degree projection of field of view,” in *International Conference on Computer Vision Workshops*, 2021, pp. 1174–1183.
- [141] P. Mirowski, A. Banki-Horvath, K. Anderson, D. Teplyashin, K. M. Hermann, M. Malinowski, M. K. Grimes, K. Simonyan, K. Kavukcuoglu, A. Zisserman *et al.*, “The streetlearn environment and dataset,” *arXiv preprint arXiv:1903.01292*, 2019.
- [142] K. Liao, C. Lin, and Y. Zhao, “A deep ordinal distortion estimation approach for distortion rectification,” *IEEE Transactions on Image Processing*, vol. 30, pp. 3362–3375, 2021.
- [143] K. Zhao, C. Lin, K. Liao, S. Yang, and Y. Zhao, “Revisiting radial distortion rectification in polar-coordinates: A new and efficient learning perspective,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 6, pp. 3552–3560, 2021.
- [144] A. Eichenseer and A. Kaup, “A data set providing synthetic and real-world fisheye video sequences,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016, pp. 1541–1545.
- [145] S. Yang, C. Lin, K. Liao, C. Zhang, and Y. Zhao, “Progressively complementary network for fisheye image rectification using appearance flow,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6348–6357.
- [146] Y. Zhao, X. Huang, and Z. Zhang, “Deep lucas-kanade homography for multimodal image alignment,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15950–15959.
- [147] R. Shao, G. Wu, Y. Zhou, Y. Fu, L. Fang, and Y. Liu, “Localtrans: A multiscale local transformer network for cross-resolution homography estimation,” in *International Conference on Computer Vision*, 2021, pp. 14890–14899.
- [148] Y. Chen, G. Wang, P. An, Z. You, and X. Huang, “Fast and accurate homography estimation using extendable compression network,” in *International Conference on Image Processing*, 2021, pp. 1024–1028.
- [149] L. Nie, C. Lin, K. Liao, S. Liu, and Y. Zhao, “Unsupervised deep image stitching: Reconstructing stitched features to images,” *IEEE Transactions on Image Processing*, vol. 30, pp. 6184–6197, 2021.
- [150] S. Garg, D. P. Mohanty, S. P. Thota, and S. Moharana, “A simple approach to image tilt correction with self-attention mobilenet for smartphones,” *arXiv preprint arXiv:2111.00398*, 2021.
- [151] K. Chen, N. Snavely, and A. Makadia, “Wide-baseline relative camera pose estimation with directional learning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3258–3268.
- [152] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niessner, M. Savva, S. Song, A. Zeng, and Y. Zhang, “Matterport3d: Learning from rgbd data in indoor environments,” *arXiv preprint arXiv:1709.06158*, 2017.
- [153] Z. Zhong, Y. Zheng, and I. Sato, “Towards rolling shutter correction and deblurring in dynamic scenes,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9219–9228.
- [154] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, “Fast and accurate image super-resolution with deep laplacian pyramid networks,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 11, pp. 2599–2613, 2018.
- [155] X. Lv, B. Wang, Z. Dou, D. Ye, and S. Wang, “Lccnet: Lidar and camera self-calibration using cost volume network,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2894–2901.
- [156] X. Lv, S. Wang, and D. Ye, “Cfnet: Lidar-camera registration using calibration flow network,” *Sensors*, vol. 21, no. 23, p. 8112, 2021.
- [157] Y. Liao, J. Xie, and A. Geiger, “Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.
- [158] B. Fan and Y. Dai, “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*, 2021, pp. 4228–4237.
- [159] B. Fan, Y. Dai, and M. He, “Sunet: symmetric undistortion network for rolling shutter correction,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 4541–4550.
- [160] Z. Liu, H. Tang, S. Zhu, and S. Han, “Semalign: Annotation-free camera-lidar calibration with semantic alignment loss,” in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2021, pp. 8845–8851.
- [161] M. Burri, J. Nikolic, P. Gohl, T. Schneider, J. Rehder, S. Omari, M. W. Achtelik, and R. Siegwart, “The euroc micro aerial vehicle datasets,” *The International Journal of Robotics Research*, vol. 35, no. 10, pp. 1157–1163, 2016.
- [162] M. Schönbein, T. Strauß, and A. Geiger, “Calibrating and centering quasi-central catadioptric cameras,” in *IEEE International Conference on Robotics and Automation*, 2014, pp. 4443–4450.
- [163] T. H. Butt and M. Taj, “Camera calibration through camera projection loss,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022, pp. 2649–2653.
- [164] X. Li, F. Flohr, Y. Yang, H. Xiong, M. Braun, S. Pan, K. Li, and D. M. Gavrila, “A new benchmark for vision-based cyclist detection,” in *IEEE Intelligent Vehicles Symposium*, 2016, pp. 1028–1033.
- [165] S.-Y. Cao, J. Hu, Z. Sheng, and H.-L. Shen, “Iterative deep homography estimation,” *arXiv preprint arXiv:2203.15982*, 2022.
- [166] M. Cao, Z. Zhong, J. Wang, Y. Zheng, and Y. Yang, “Learning adaptive warping for real-world rolling shutter correction,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17785–17793.
- [167] T. Do, O. Miksik, J. DeGol, H. S. Park, and S. N. Sinha, “Learning to detect scene landmarks for camera localization,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11132–11142.
- [168] T. Do, K. Vuong, S. I. Roumeliotis, and H. S. Park, “Surface normal estimation of tilted images via spatial rectifier,” in *European Conference on Computer Vision*, 2020, pp. 265–280.
- [169] J. Shotton, B. Glocker, C. Zach, S. Izadi, A. Criminisi, and A. Fitzgibbon, “Scene coordinate regression forests for camera re-localization in rgbd images,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 2930–2937.
- [170] C. M. Parameshwara, G. Hari, C. Fermüller, N. J. Sanket, and Y. Aloimonos, “Diffposenet: Direct differentiable camera pose estimation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 6845–6854.
- [171] W. Wang, D. Zhu, X. Wang, Y. Hu, Y. Qiu, C. Wang, Y. Hu, A. Kapoor, and S. Scherer, “Tartanair: A dataset to push the limits of visual slam,” in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2020, pp. 4909–4916.
- [172] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, “A benchmark for the evaluation of rgbd slam systems,” in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2012, pp. 573–580.
- [173] B. J. Pijnacker Hordijk, K. Y. Schepers, and G. C. De Croon, “Vertical landing for micro air vehicles using event-based optical flow,” *Journal of Field Robotics*, vol. 35, no. 1, pp. 69–90, 2018.
- [174] L. Yang, R. Shrestha, W. Li, S. Liu, G. Zhang, Z. Cui, and P. Tan, “Scenesqueezor: Learning to compress scene for camera relocalization,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 8259–8268.
- [175] W. Maddern, G. Pascoe, C. Linegar, and P. Newman, “1 year, 1000 km: The oxford robotcar dataset,” *The International Journal of Robotics Research*, vol. 36, no. 1, pp. 3–15, 2017.
- [176] G. Poniatkin, Y. Labb  , B. Russell, M. Aubry, and J. Sivic, “Focal length and object pose estimation via render and compare,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3825–3834.
- [177] X. Sun, J. Wu, X. Zhang, Z. Zhang, C. Zhang, T. Xue, J. B. Tenenbaum, and W. T. Freeman, “Pix3d: Dataset and methods for single-image 3d shape modeling,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2974–2983.
- [178] Y. Wang, X. Tan, Y. Yang, X. Liu, E. Ding, F. Zhou, and L. S. Davis, “3d pose estimation for fine-grained object categories,” in *European Conference on Computer Vision Workshops*, 2018.
- [179] X. Jing, X. Ding, R. Xiong, H. Deng, and Y. Wang, “Dxq-net: Differentiable lidar-camera extrinsic calibration using quality-aware flow,” *arXiv preprint arXiv:2203.09385*, 2022.
- [180] Y. Zhang, X. Zhao, and D. Qian, “Learning-based framework for camera calibration with distortion correction and high precision feature detection,” *arXiv preprint arXiv:2202.00158*, 2022.
- [181] Y. Sun, J. Li, Y. Wang, X. Xu, X. Yang, and Z. Sun, “Atop: An attention-to-optimization approach for automatic lidar-camera calibration via cross-modal object matching,” *IEEE Transactions on Intelligent Vehicles*, 2022.
- [182] G. Wang, J. Qiu, Y. Guo, and H. Wang, “Fusionnet: Coarse-to-fine extrinsic calibration network of lidar and camera with hierarchical point-pixel fusion,” in *International Conference on Robotics and Automation*, 2022, pp. 8964–8970.
- [183] C. Ye, H. Pan, and H. Gao, “Keypoint-based lidar-camera online calibration with robust geometric network,” *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–11, 2021.
- [184] N. Wakai, S. Sato, Y. Ishii, and T. Yamashita, “Rethinking generic camera models for deep single image camera calibration to recover rotation and fisheye distortion,” in *European Conference on Computer Vision*, vol. 13678, 2022, pp. 679–698.
- [185] S.-H. Chang, C.-Y. Chiu, C.-S. Chang, K.-W. Chen, C.-Y. Yao, R.-R. Lee, and H.-K. Chu, “Generating 360 outdoor panorama dataset with reliable sun position estimation,” in *SIGGRAPH Asia*, 2018, pp. 1–2.
- [186] C. Wu, “Towards linear-time incremental structure from motion,” in *International Conference on 3D Vision*, 2013, pp. 127–134.
- [187] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan *et al.*, “Searching for mobilenetv3,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1314–1324.
- [188] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, “Pointnet: Deep learning on point sets for 3d classification and segmentation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 652–660.
- [189] Y. Li, R. Bu, M. Sun, W. Wu, X. Di, and B. Chen, “Pointcnn: Convolution on x-transformed points,” *Advances in Neural Information Processing Systems*, vol. 31, 2018.
- [190] Z. Wang, S. Wu, W. Xie, M. Chen, and V. A. Prisacariu, “Nerf: Neural radiance fields without known camera parameters,” *arXiv preprint arXiv:2102.07064*, 2021.
- [191] Y. Jeong, S. Ahn, C. Choy, A. Anandkumar, M. Cho, and J. Park, “Self-calibrating neural radiance fields,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 5846–5854.
- [192] P. Truong, M.-J. Rakotosaona, F. Manhardt, and F. Tombari, “Sparf: Neural radiance fields from sparse and noisy poses,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 4190–4200.
- [193] K. Wang, Z. Yan, H. Tian, Z. Zhang, X. Li, J. Li, and J. Yang, “Altnerf: Learning robust neural radiance field via alternating depth-pose optimization,” *arXiv preprint arXiv:2308.10001*, 2023.
- [194] W. Bian, Z. Wang, K. Li, J.-W. Bian, and V. A. Prisacariu, “Nooperf: Optimising neural radiance field with no pose prior,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 4160–4169.
- [195] Q. Meng, A. Chen, H. Luo, M. Wu, H. Su, L. Xu, X. He, and J. Yu, “Gnerf: Gan-based neural radiance field without posed camera,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 6351–6361.
- [196] S.-F. Chng, S. Ramasinghe, J. Sherrah, and S. Lucey, “Gaussian activated neural radiance fields for high fidelity reconstruction and pose estimation,” in *European Conference on Computer Vision*, 2022, pp. 264–280.
- [197] Y. Xia, H. Tang, R. Timofte, and L. Van Gool, “Sinerf: Sinusoidal neural radiance fields for joint pose estimation and scene reconstruction,” *arXiv preprint arXiv:2210.04553*, 2022.
- [198] C.-H. Lin, W.-C. Ma, A. Torralba, and S. Lucey, “Barf: Bundle-adjusting neural radiance fields,” in *International Conference on Computer Vision*, 2021, pp. 5741–5751.
- [199] L. Yen-Chen, P. Florence, J. T. Barron, A. Rodriguez, P. Isola, and T.-Y. Lin, “iNeRF: Inverting neural radiance fields for pose estimation,” in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2021.
- [200] T. M  ller, A. Evans, C. Schied, and A. Keller, “Instant neural graphics primitives with a multiresolution hash encoding,” *ACM Transactions on Graphics*, vol. 41, no. 4, pp. 1–15, 2022.

- [201] M. Tancik, E. Weber, E. Ng, R. Li, B. Yi, T. Wang, A. Kristoffersen, J. Austin, K. Salahi, A. Ahuja *et al.*, "Nerfstudio: A modular framework for neural radiance field development," in *ACM SIGGRAPH Conference Proceedings*, 2023, pp. 1–12.
- [202] W. Xian, A. Božić, N. Snavely, and C. Lassner, "Neural lens modeling," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 8435–8445.
- [203] S. Zhu, A. Kumar, M. Hu, and X. Liu, "Tame a wild camera: In-the-wild monocular camera calibration," *arXiv preprint arXiv:2306.10988*, 2023.
- [204] L. Jin, J. Zhang, Y. Hold-Geoffroy, O. Wang, K. Blackburn-Matzen, M. Sticha, and D. F. Fouhey, "Perspective fields for single image camera calibration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 17 307–17 316.
- [205] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz, "Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.
- [206] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge University Press, 2003.
- [207] B. D. Lucas, T. Kanade *et al.*, *An iterative image registration technique with an application to stereo vision*. Vancouver, 1981, vol. 81.
- [208] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "Shufflenet v2: Practical guidelines for efficient cnn architecture design," in *European conference on computer vision*, 2018, pp. 116–131.
- [209] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.
- [210] L. Nie, C. Lin, K. Liao, and Y. Zhao, "Learning edge-preserved image stitching from multi-scale deep homography," *Neurocomputing*, vol. 491, pp. 533–543, 2022.
- [211] S. Baker and I. Matthews, "Lucas-kanade 20 years on: A unifying framework," *International Journal of Computer Vision*, vol. 56, no. 3, pp. 221–255, 2004.
- [212] J. Nocedal and S. J. Wright, *Numerical optimization*. Springer, 1999.
- [213] Z. Teed and J. Deng, "Raft: Recurrent all-pairs field transforms for optical flow," in *European Conference on Computer Vision*, 2020, pp. 402–419.
- [214] Y. Li, W. Pei, and Z. He, "Ssorn: Self-supervised outlier removal network for robust homography estimation," *arXiv preprint arXiv:2208.14093*, 2022.
- [215] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [216] A. Handa, M. Bloesch, V. Pătrăucean, S. Stent, J. McCormac, and A. Davison, "gvnn: Neural network library for geometric computer vision," in *European Conference on Computer Vision*. Springer, 2016, pp. 67–82.
- [217] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [218] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.
- [219] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "Pointpillars: Fast encoders for object detection from point clouds," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 697–12 705.
- [220] R. Poli, J. Kennedy, and T. Blackwell, "Particle swarm optimization," *Swarm Intelligence*, vol. 1, no. 1, pp. 33–57, 2007.
- [221] S. Gould, R. Hartley, and D. Campbell, "Deep declarative networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 8, pp. 3988–4004, 2021.
- [222] B. Xiao, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," in *European Conference on Computer Vision*, 2018, pp. 466–481.
- [223] X. Huang, P. Wang, X. Cheng, D. Zhou, Q. Geng, and R. Yang, "The apolloscape open dataset for autonomous driving and its application," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 10, pp. 2702–2719, 2019.
- [224] H. Yu, Y. Luo, M. Shu, Y. Huo, Z. Yang, Y. Shi, Z. Guo, H. Li, X. Hu, J. Yuan *et al.*, "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*, 2022, pp. 21 361–21 370.
- [225] J. Kang and N. L. Doh, "Automatic targetless camera–LIDAR calibration by aligning edge with Gaussian mixture model," *Journal of Field Robotics*, vol. 37, no. 1, pp. 158–179, 2020.
- [226] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liou, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 621–11 631.
- [227] J. Mao, M. Niu, C. Jiang, X. Liang, Y. Li, C. Ye, W. Zhang, Z. Li, J. Yu, C. Xu *et al.*, "One million scenes for autonomous driving: Once dataset," 2021.
- [228] C. Tang and P. Tan, "Ba-net: Dense bundle adjustment network," *arXiv preprint arXiv:1806.04807*, 2018.
- [229] Z. Teed and J. Deng, "Deepv2d: Video to depth with differentiable structure from motion," *arXiv preprint arXiv:1812.04605*, 2018.
- [230] X. Wei, Y. Zhang, Z. Li, Y. Fu, and X. Xue, "Deepsfm: Structure from motion via deep bundle adjustment," in *European Conference on Computer Vision*, 2020, pp. 230–247.
- [231] X. Gu, W. Yuan, Z. Dai, S. Zhu, C. Tang, Z. Dong, and P. Tan, "Dro: Deep recurrent optimizer for video to depth," *IEEE Robotics and Automation Letters*, vol. 8, no. 5, pp. 2844–2851, 2023.
- [232] A. Hagemann, M. Knorr, and C. Stiller, "Deep geometry-aware camera self-calibration from video," in *International Conference on Computer Vision*, 2023, pp. 3438–3448.
- [233] K. He, R. Girshick, and P. Dollár, "Rethinking imagenet pre-training," in *International Conference on Computer Vision*, 2019, pp. 4918–4927.

Deep Learning for Camera Calibration and Beyond: A Survey -Supplementary Material-

Kang Liao, Lang Nie, Shujuan Huang, Chunyu Lin, Jing Zhang, Yao Zhao, *Fellow, IEEE*, Moncef Gabbouj, *Fellow, IEEE*, Dacheng Tao, *Fellow, IEEE*

1 DEVELOPMENT RECAP

1.1 Milestone

A concrete milestone from 2015 to 2022 of deep learning-based camera calibration is shown in Figure 1, spanning the main deep learning era. We classify all literature based on the uncalibrated camera model and its extended applications: standard model, distortion model, cross-view model, and cross-sensor model.

1.2 Statistic Analysis

As we can observe in Figure 2, the number of learning-based camera calibrations has grown since 2015 and boomed since 2019. And the learning targets are extended from the simple and pure parameters to complicated and hybrid parameters, driven by larger datasets, more reasonable learning strategies, more explicit learning representations, and more solid network architectures, etc.

The data analysis of different learning strategies used in learning-based camera calibration is also shown in Figure 2. From the statistics, six strategies have been investigated, in which supervised learning accounts for the largest majority (more than 90%). Considering the expensive labeling works, some recent research explores liberating the training demand for camera parameters using semi-supervised learning, weakly-supervised learning, unsupervised learning, and self-supervised learning. Reinforcement learning also has been exploited to dynamically address the camera calibration problem.

2 CAMERA MODEL

Researchers utilize mathematical formulas to establish camera models that describe the imaging process from a point in 3D world coordinates to its projection on a 2D image plane. Different cameras and systems correspond to different types of parametric models. In this section, we first provide a detailed formulation of the basic pinhole camera model. Then, we review more complex and useful camera models, as well as extended models studied in recent literature, to meet the advanced development of cameras and academic/industrial demands.

2.1 Pinhole Camera Model

The most popular and commonly applied camera model in computer vision is the pinhole camera model. It can be regarded as a geometrically accurate first-order approximation of the traditional camera. A pinhole camera has one single effective perspective because the pinhole aperture is thought to be an infinitesimal point through which all projection lines must pass.

Using a mathematical formulation, the camera model depicts the imaging process from a point in the 3D world coordinate to its projection on the 2D image plane. Assuming the homogeneous coordinates $\mathbf{P}_w = [X, Y, Z, 1]^T \in \mathbb{R}^{4 \times 1}$ and $\mathbf{P}_i = [u, v, 1]^T \in \mathbb{R}^{3 \times 1}$ denote a point in the 3D world coordinate and its corresponding point on a 2D image plane, respectively. Then, a camera model can be described by a projection mapping $M \in \mathbb{R}^{3 \times 4}$ between \mathbf{P}_w and \mathbf{P}_i :

$$\mathbf{P}_i = M\mathbf{P}_w, \quad (1)$$

where the projection can be further formed by:

$$\mathbf{P}_c = [\mathbf{R} | \mathbf{t}]\mathbf{P}_w, \quad (2)$$

where $\mathbf{P}_c = [x_c, y_c, z_c]^T \in \mathbb{R}^{3 \times 1}$ denotes a transformed point in the camera coordinate using a 3×3 rotation \mathbf{R} and a 3-dimension translation \mathbf{t} . The 3×4 matrix $[\mathbf{R} | \mathbf{t}]$ is generally named as the extrinsic camera matrix, in which the camera rotation can be further parameterized by three angles: yaw φ , pitch θ , and roll ψ . Subsequently, the point \mathbf{P}_c is projected onto a surface. This surface is represented by the pinhole camera model as a plane $z = 1$, and the normalized coordinate of the point in camera coordinate is expressed by $[x_n, y_n]^T = [\frac{x_c}{z_c}, \frac{y_c}{z_c}]^T$.

Finally, the point on the normalized plane is projected onto the image plane, obtaining a pixel \mathbf{P}_i by:

$$\mathbf{P}_i = K[x_n, y_n, 1]^T, \quad (3)$$

where $K \in \mathbb{R}^{3 \times 3}$ is an intrinsic camera matrix, which consists of various camera intrinsic parameters such as the focal length, skew coefficient, and image center:

$$K = \begin{bmatrix} f_x m_u & s & c_u \\ 0 & f_y m_v & c_v \\ 0 & 0 & 1 \end{bmatrix}, \quad (4)$$

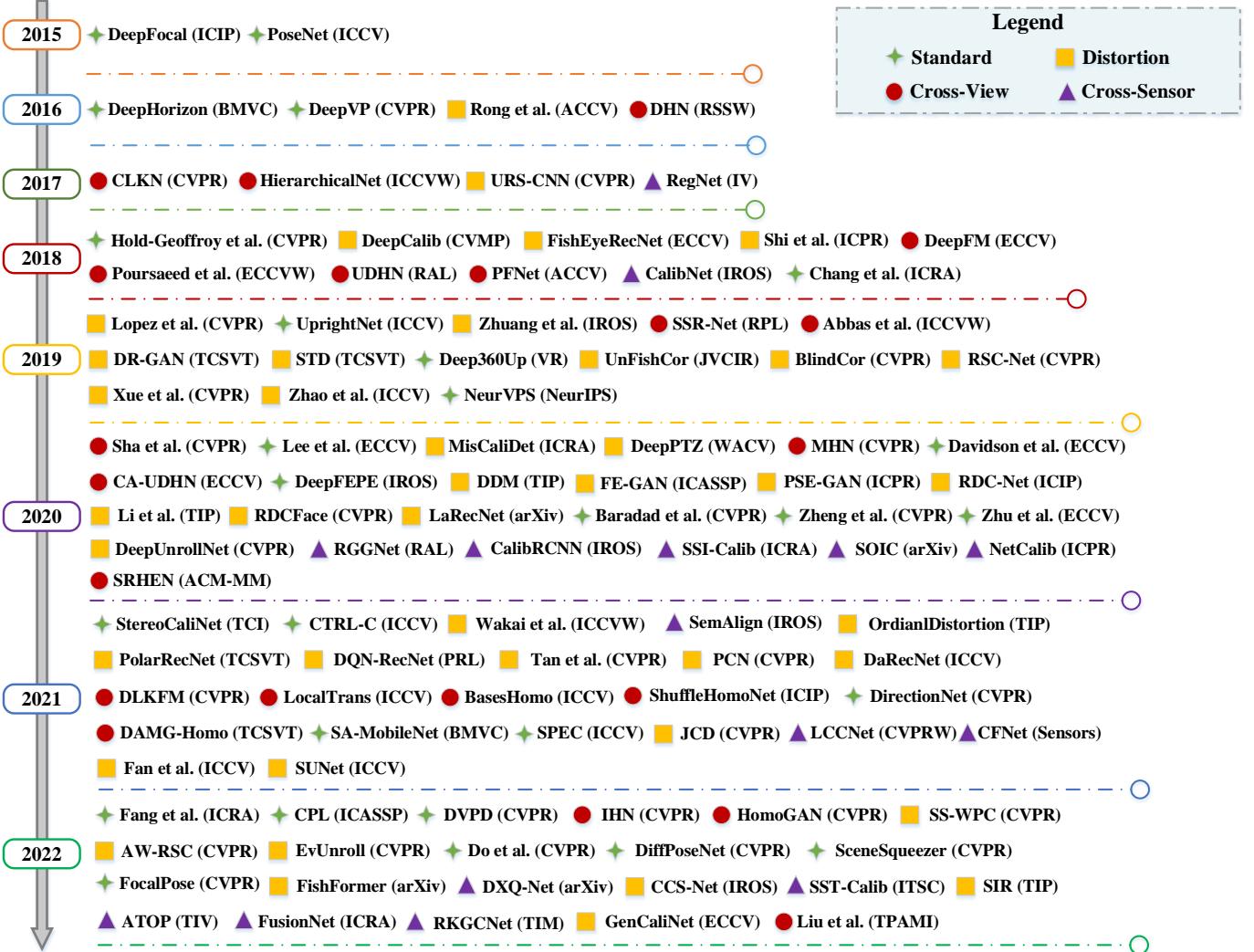


Fig. 1. A concise milestone of deep learning-based camera calibration methods. We classify all methods based on the uncalibrated camera model and its extended applications: standard model, distortion model, cross-view model, and cross-sensor model. **Standard model**: DeepFocal [1], PoseNet [2], DeepHorizon [3], DeepVP [4], Chang et al. [5], UprightNet [6], Lee et al. [7], NeurVPS [8], Deep360Up [9], Davidson et al. [10], DeepFEPE [11], Baradad et al. [12], Zheng et al. [13], Zhu et al. [14], StereoCalibNet [15], SA-MobileNet [16], Fang et al. [17], CPL [18], FocalPose [19], DirectionNet [20], DVPD [21], Do et al. [22], CTRL-C [23], SPEC [24], DiffPoseNet [25], SceneSqueezer [26]. **Distortion model**: Rong et al. [27], Hold-Geoffroy et al. [28], DeepCalib [29], URS-CNN [30], FishEyeRecNet [31], Shi et al. [32], DR-GAN [33], STD [34], UnFishCor [35], BlindCor [36], RSC-Net [37], Xue et al. [38], Zhuang et al. [39], Lopez et al. [40], Zhao et al. [41], DDM [42], MisCalibDet [43], DeepPTZ [44], FE-GAN [45], PSE-GAN [46], RDC-Net [47], Li et al. [48], RDCFace [49], LaRecNet [50], DeepUnrollNet [51], OrdianlDistortion [52], PolarRecNet [53], DQN-RecNet [54], Tan et al. [55], PCN [56], SIR [57], DaRecNet [58], Wakai et al. [59], GenCalibNet [60], JCD [61], SS-WPC [62], AW-RSC [63], EvUnroll [64], Fan et al. [65], SUNet [66], CCS-Net [67], FishFormer [68]. **Cross-View model**: DHN [69], CLKN [70], HierarchicalNet [71], DeepFM [72], Poursaeed et al. [73], UDHN [74], PFNet [75], SRHEN [76], SSR-Net [77], Abbas et al. [78], Sha et al. [79], MHN [80], CA-UDHN [81], DLKFM [82], LocalTrans [83], BasesHomo [84], ShuffleHomoNet [85], DAMG-Homo [86], IHN [87], HomoGAN [88], Liu et al. [89]. **Cross-Sensor model**: RegNet [90], CalibNet [91], RGGNet [92], SSI-Calib [93], SOIC [94], CalibRCNN [95], NetCalib [96], LCCNet [97], CFNet [98], SemAlign [99], DXQ-Net [100], SST-Calib [101], ATOP [102], FusionNet [103], RGKCNNet [104].

where f_x and f_y are the focal lengths at X-axis and Y-axis of the camera, respectively. Generally, for most cameras, $f_x = f_y$, and they are unified to f . m_u and m_v are the number of pixels per unit distance, in which $m_u = m_v$, if the image has square pixels. s is the skew coefficient. A CCD sensor's pixels might not be precisely square, which would cause a slight distortion in the X or Y axes. The number of pixels on the CCD sensor per unit length in each direction is known as the skew coefficient. It would become 0 when X-axis and Y-axis are perpendicular to each other. $[c_u, c_v]^T$ is the coordinate of the image center. According to previous works and factory design, the intrinsic parameters can be refined by $s = 0$, $m_u = m_v$ and focal length in the pixel

unit, then Eq. (3) can be reformulated as:

$$\mathbf{P}_i = \begin{bmatrix} f_x & 0 & c_u \\ 0 & f_y & c_v \\ 0 & 0 & 1 \end{bmatrix} [x_n, y_n, 1]^T. \quad (5)$$

In addition to numerical camera parameters, some geometric representations can provide useful clues for camera calibration, such as vanishing points and horizon lines. These representations establish clear relationships between image features and calibration objectives, which can alleviate the difficulty of learning conventional and implicit camera parameters.

Lines and points are both represented as three-dimensional vectors in homogeneous coordinates. The def-

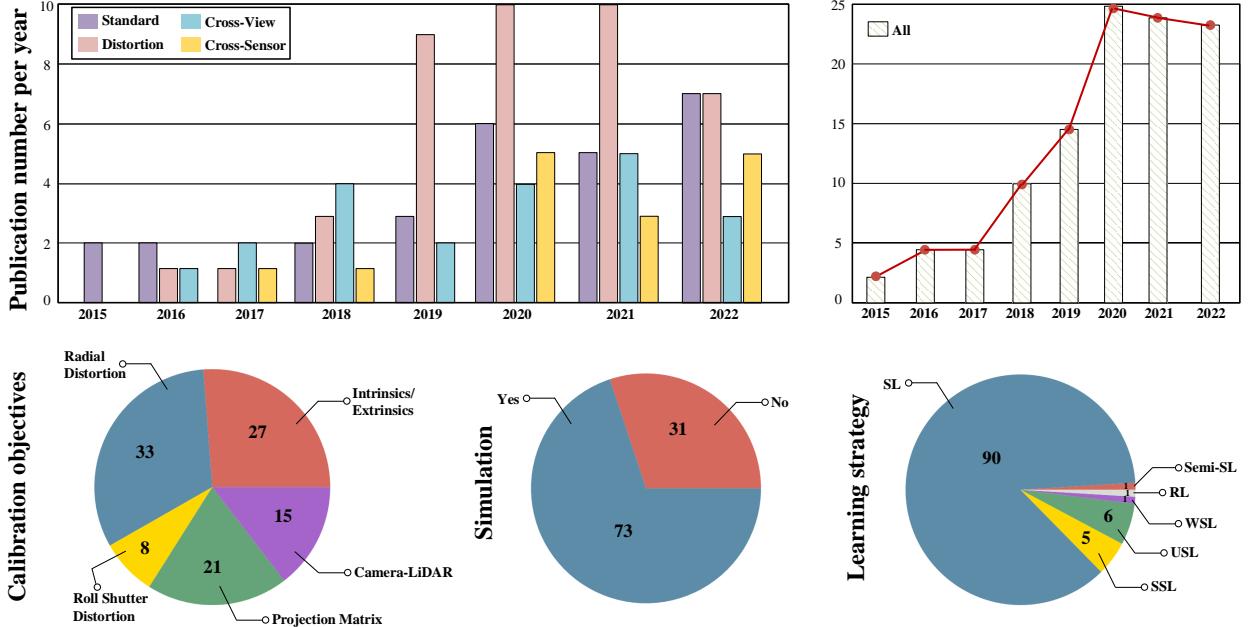


Fig. 2. A statistic analysis of deep learning-based camera calibration methods. To be specific, we summarize all literature based on the number of publications per year, calibration objectives, simulation of the dataset, and learning strategy.

initions for computing the line \mathbf{l} that connects two points and the point \mathbf{p} at the intersection of two lines can be given by:

$$\mathbf{l} = \frac{\mathbf{p}_1 \times \mathbf{p}_2}{\|\mathbf{p}_1 \times \mathbf{p}_2\|} \quad \mathbf{p} = \frac{\mathbf{l}_1 \times \mathbf{l}_2}{\|\mathbf{l}_1 \times \mathbf{l}_2\|} \quad (6)$$

There are two parameterizations of the horizon line: slope-offset (θ, ρ) and left-right (l, r). Assuming that the viewing orientation is down the negative z -axis, with the positive x -direction to the right, and the positive y -direction to the up. As a result, the world viewing direction of the camera can be described by $R_c^T [0, 0, -1]^T$. For the world vector $[0, 1, 0]^T$ points in the zenith direction, a set of points \mathbf{p} can represent the horizon line:

$$\mathbf{p}^T K^{-T} R [0, 1, 0]^T = 0. \quad (7)$$

As mentioned in Barnard [105], the normalized line direction vector \mathbf{d} can be formulated for the Gaussian sphere representation of a vanishing point \mathbf{v} . In particular, supposed a 3D ray is described by $\mathbf{o} + \lambda \mathbf{d}$, where \mathbf{o} and \mathbf{d} are its origin and unit direction vector, respectively. Then, the vanishing point can be represented by $\lambda \rightarrow \infty$, in which the image coordinate is formed by $\mathbf{v} = [v_x, v_y]^T := \lim_{\lambda \rightarrow \infty} [p_x, p_y]^T \in \mathbb{R}^2$. Thus, the 3D direction of a line based on its vanishing point can be calculated by:

$$\mathbf{d} = [v_x - c_x \quad v_y - c_y \quad f]^T \in \mathbb{R}^3. \quad (8)$$

By using \mathbf{d} rather than \mathbf{v} , the degraded situations where \mathbf{d} is parallel to the image plane are eliminated. Additionally, it provides a natural measurement for determining the separation between two vanishing points.

2.2 Wide-angle Camera Model

The perspective projection model, given a typical pinhole camera with focal length f , can be expressed as:

$$r = f \tan \theta, \quad (9)$$

where r denotes the projection distance between the principal point and the points in the image. θ denotes the angle between the incident ray and the optical axis of the camera. It is straightforward to determine that θ should be less than 90° . Without a projection point on the image plane, the incoming ray will not cross with the image plane and the pinhole camera will not be able to view anything behind. Because of their restricted field of view (FoV), most cameras cannot see all of the points in the 3D environment at the same time.

Due to the wide FoV, wide-angle cameras are increasingly widely used in computer vision and robotics tasks such as navigation, localization, and tracking. Specifically, an extra wide-angle lens called a fisheye camera is used to create a broad, hemispherical, or panoramic image. Fisheye lenses employ a specific mapping to produce convex and non-rectilinear images as opposed to images with straight lines of perspective (rectilinear images). However, the wide-angle camera violates the pinhole camera assumption and the captured image suffers from geometric distortions.

Geometric distortion induced by wide-angle cameras can generally be classified into radial distortion and tangential distortion (de-centering distortion). Radial distortion is the primary distortion in central single-view camera systems, exhibiting circular symmetry with respect to the distortion center. This distortion results in points on the image plane being moved away from their ideal location under the perspective camera model along the radial axis from the distortion center. Radial distortion models can be formulated as nonlinear functions of the radial distance [106]. On the other hand, tangential distortion occurs when the lens and image plane are not parallel. Tangential distortion, also known as de-centering distortion, is primarily caused by the lens assembly not being centered over and parallel to the image plane. Unlike radial distortion, tangential distortion has a

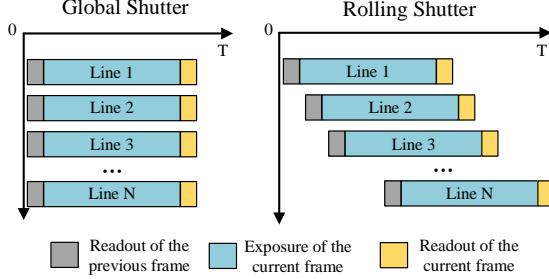


Fig. 3. Comparison of the mechanism of global shutter camera and rolling shutter camera.

geometric impact that is not solely along the radial axis, and can also cause rotation and skewing of the image plane with respect to the distance from the image center. The camera model with radial distortion and tangential distortion can be parameterized by:

$$\begin{cases} x_r = x_d + \bar{x}(k_1 r_d^2 + k_2 r_d^4 + k_3 r_d^6 + \dots) \\ y_r = y_d + \bar{y}(k_1 r_d^2 + k_2 r_d^4 + k_3 r_d^6 + \dots) \end{cases}, \quad (10)$$

where $\bar{x} = x_d - c_x$ and $\bar{y} = y_d - c_y$. $K = (k_1, k_2, k_3, \dots)$ and $P = (p_1, p_2, p_3, \dots)$ are the radial distortion parameters and decentering distortion parameters, respectively. r_d describes the radial distance from an image point to the distortion center (c_x, c_y) . Such an equation represents the mapping from a point $[x_d, y_d]^\top$ in the image captured by the wide-angle camera to that in the rectified image without the geometric distortion $[x_r, y_r]^\top$.

Previous works demonstrate that tangential distortion is basically insignificant and can be neglected. Moreover, as we surveyed, all learning-based camera calibration methods only consider the radial distortion for calibrating the wide-angle camera. To this end, Eq.10 can be simplified by a Taylor expansion:

$$\begin{cases} x_r = x_d(k_1 r_d^2 + k_2 r_d^4 + k_3 r_d^6 + \dots) \\ y_r = y_d(k_1 r_d^2 + k_2 r_d^4 + k_3 r_d^6 + \dots) \end{cases}. \quad (11)$$

This equation is known as the even-order polynomial model, which can also be expressed as an odd-order polynomial model by shifting the power. However, according to Wang et al. [107], while the polynomial model is suitable for small distortions, it requires an unreasonably high number of non-zero distortion parameters for severe distortions. As an alternative, Fitzgibbon et al. [108] proposed a division model that more accurately approximates the genuine undistortion function of a common camera. For significant distortion, the division model is preferred over the polynomial model because it requires fewer terms:

$$\begin{cases} x_r = \frac{x_d}{k_1 r_d^2 + k_2 r_d^4 + k_3 r_d^6 + \dots} \\ y_r = \frac{y_d}{k_1 r_d^2 + k_2 r_d^4 + k_3 r_d^6 + \dots} \end{cases}. \quad (12)$$

Some classical works demonstrate the single-parameter division model (only with distortion parameter k_1 in Eq.12) seems to be sufficient for most wide-angle cameras, which has been widely applied in learning-based wide-angle camera calibration [27], [29], [36], [54].

2.3 Rolling Shutter Camera Model

Due to the compact design, low price, and high frame rate, numerous consumer cameras, including webcams and mobile phones, employ CMOS (complementary metal–oxide–semiconductor) sensors. But they are restricted to using rolling shutter (RS) devices. With a consistent time delay between each row, RS exposes the sensor array row by row from top to bottom, as opposed to global shutter (GS) based on CCD sensors, which simultaneously read out all rows of the sensor array. If the RS camera is moving while capturing the image, various distortions, such as skew, smear, or wobble, will break the reality of the original scene, which deviates from the pinhole camera paradigm. The unknown camera movements during the capturing process induce the so-called RS effects (also known as the jelly effect). In other words, an RS image is a row-by-row combination of GS images taken by a virtual moving GS camera throughout the camera readout time. The comparison of the RS camera and GS camera is shown in Figure 3.

The RS camera can be regarded as a high-frequency sensor that produces sparse spatial information with rich temporal coverage conveyed by distortions [109]. Modeling the RS camera faces a common challenge of estimating the transformation between RS and GS images. Assume a 3D latent space-time volume captures the desired scene across the desired time period $[0, t_0]$ and creates a virtual GS image \mathbf{I}^{GS} . We suppose the readout direction is from top to bottom, and then the row-by-row readout RS imaging \mathbf{I}^{RS} can be expressed by:

$$\mathbf{I}^{RS} = \sum_{y=1}^H M(\mathbf{I}_{t_r}^{GS}, y), \quad (13)$$

where H is the height of the RS image (total number of rows) and y indicates the vertical coordinate. $M(\cdot, \cdot)$ masks a specific row in the GS image, in which t_r represents the readout (offset) time for each row of RS.

On the other hand, by warping the RS features backward with an estimated displacement field, the GS image can be formulated by:

$$\mathbf{I}^{GS}(\mathbf{x}) = \mathbf{I}^{RS}(\mathbf{x} + \mathbf{F}_{GS \rightarrow RS}(\mathbf{x})), \quad (14)$$

where $\mathbf{F}_{GS \rightarrow RS} \in \mathbb{R}^2$ denotes the displacement field of the pixel \mathbf{x} from the GS image to the RS image.

The above formulations describe the rolling shutter camera model under a short exposure scenario. When the exposure time of the camera increases, the motion blur effects occur in the captured image, jointly with the RS distortion:

$$\mathbf{I}_{(t)}^{RS'}[i] = \frac{1}{T} \int_{t-t_h+it_r-T/2}^{t-t_h+it_r+T/2} \mathbf{I}_{(t-t_h+it_r)}^{GS}[i] dt, \quad (15)$$

where $\mathbf{I}_{(t)}^{RS'}[i]$ denotes the i^{th} row of the RS distortion image \mathbf{I}^{RS} with the middle moment of exposure at time t . T indicates the exposure time of camera and $t_h = (H/2)t_r$.

2.4 Cross-View Camera Model

The cross-view camera model is a type of multi-view camera system used in computer vision. It involves placing two or more cameras at opposite sides of a scene to capture multiple views of the same scene. This setup enables the

creation of 3D reconstructions of the scene by triangulating corresponding points from multiple camera views. The cross-view camera model is commonly used in surveillance, robotics, and augmented reality applications, and provides a more accurate and complete representation of the scene than what can be achieved with a single camera. Alternatively, a camera with stable movement can also be regarded as a cross-view camera model.

In a cross-view camera model, the captured images can be used to calculate the fundamental matrix and homography matrix, which are essential tools for 3D reconstruction, image rectification, and camera calibration.

Fundamental Matrix Geometric relationships between the 3D points and their projections onto the 2D plane impose constraints on the image points when two cameras capture the same 3D scene from different perspectives. This intrinsic projective geometry can be embodied by a fundamental matrix \mathbf{F} .

$$\mathbf{F} = \mathbf{K}_2^{-T} [\mathbf{t}]_{\times} \mathbf{R} \mathbf{K}_1^{-1}. \quad (16)$$

Such an equation describes the epipolar geometry, where \mathbf{K}_1 and \mathbf{K}_2 indicate the intrinsic parameters of two cameras, and \mathbf{R} and $[\mathbf{t}]_{\times}$ are the relative camera rotation and translation, respectively.

The fundamental matrix can be calculated from the correspondences of projected scene points by $q^T \mathbf{F} p = 0$, in which q and p are the matching points derived from two views. Specifically, the eight-point algorithm [110] uses 8 point correspondences and enforces the rank-2 constraint using Singular Value Decomposition (SVD), computing a matrix with the minimum Frobenius distance.

Homography Matrix Estimating a 2D homography matrix (or projection transformation) is an elemental geometric task for a pair of images that are captured from the same planar surface in a 3D scene with different perspectives. An invertible mapping from one image plane to another with eight degrees of freedom: two each for translation, rotation, scale, and lines at infinity, is known as a homography. Supposed that the homogeneous coordinates $\mathbf{x} = [u, v, 1]^T \in \mathbb{R}^{3 \times 1}$ and $\mathbf{x}' = [u', v', 1]^T \in \mathbb{R}^{3 \times 1}$ are points from two images but indicating the same point in the 3D scene, a non-singular 3×3 matrix can represent a linear transformation that maps $\mathbf{x} \Leftrightarrow \mathbf{x}'$ as a planar projective transformation or homography \mathbf{H} :

$$\begin{bmatrix} u' \\ v' \\ 1 \end{bmatrix} \sim \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} \begin{bmatrix} u \\ v \\ 1 \end{bmatrix}, \quad (17)$$

where the transformation can be simplified as $\mathbf{x}' \sim \mathbf{Hx}$. This transformation can be rewritten by two following equations:

$$u' = \frac{h_{11}u + h_{12}v + h_{13}}{h_{31}u + h_{32}v + h_{33}}, \quad v' = \frac{h_{21}u + h_{22}v + h_{23}}{h_{31}u + h_{32}v + h_{33}}. \quad (18)$$

Previous methods [69], [74] point out that the above conventional 3×3 parameterization \mathbf{H} is not desirable for training neural networks. Concretely, it is challenging to guarantee the non-singularity of \mathbf{H} due to the significant variance in the size of the members of the 3×3 homography matrix. Moreover, the rotation, translation, scale, and shear components of the homography transformation are mixed in \mathbf{H} . For instance, the submatrix $[h_{11} \ h_{12}; h_{21} \ h_{22}]$

describes the homography's rotational term and the vector $[h_{13}, h_{23}]^T$ denotes the translation transformation. Considering the rotation and shear components typically have smaller magnitudes than the translation component, it will have a negligible impact on the loss function of the component elements, leading to an imbalance training problem with a neural network. Instead, a 4-point parameterization [111] has been demonstrated to be more learning-friendly for learning-based homography estimation than the 3×3 parameterization. Supposed that the offsets of the image's vertex are $\Delta u_i = u'_i - u_i$ and $\Delta v_i = v'_i - v_i$, then the 4-point parameterization $\tilde{\mathbf{H}}$ can describe a homography by:

$$\tilde{\mathbf{H}} = \begin{pmatrix} \Delta u_1 & \Delta v_1 \\ \Delta u_2 & \Delta v_2 \\ \Delta u_3 & \Delta v_3 \\ \Delta u_4 & \Delta v_4 \end{pmatrix}. \quad (19)$$

The 4-point parameterization owns eight variables, which are equivalent to the matrix formulation of the homography. It is straightforward to transform from $\tilde{\mathbf{H}}$ to \mathbf{H} using the normalized Direct Linear Transform (DLT) [112] if the four corners' displacement is known.

2.5 Cross-Sensor Model

Modern robots are often equipped with various sensors to provide a comprehensive understanding of the environment. These sensors capture scenes using different types of representations. For autonomous cars and robotics, cameras and Light Detection and Ranging sensors (LiDAR) are commonly used for vision tasks. The 3D LiDAR records long-range spatial data as sparse point clouds, while the camera captures texturally dense 2D color RGB images. Combining these sensors can facilitate 3D reconstruction and provide precise and robust perception for the robots, overcoming the limitations of individual sensors.

However, collision and vibration problems can occur when using different sensors in a robot or system. Additionally, the 3D point clouds cannot be effectively projected onto a 2D image without accurate extrinsic parameters, making it difficult to reliably correlate pixels in an image with depth information. Therefore, it is crucial to precisely calibrate the 2D-3D matching correspondences between pairs of temporally synchronized camera and LiDAR data.

The appropriate extrinsic calibration of the transformation (*i.e.*, rotation and translation) between the camera and LiDAR in 6-DoF is a key condition for data fusion. To be more specific, 3D LiDAR point cloud $PC = [X, Y, Z] \in \mathbb{R}^3$ can be projected onto the image plane by transforming it into the camera coordinate using the extrinsic matrix T between the camera and LiDAR as well as camera intrinsic K . The inverse depth and the projected 2D coordinates can be represented as $d = 1/Z$ and $p = [u, v] \in \mathbb{R}^2$, respectively. Then, the camera-LiDAR model can be described by:

$$\begin{bmatrix} u \\ v \\ d \end{bmatrix} = \begin{bmatrix} f_x(\hat{X}/\hat{Z}) + c_x \\ f_y(\hat{Y}/\hat{Z}) + c_y \\ 1/\hat{Z} \end{bmatrix}, \quad (20)$$

where (f_x, f_y) and (c_x, c_y) indicate the focal lengths and the image center as listed in Eq. 4. $[\hat{X}, \hat{Y}, \hat{Z}]$ is the transformed point cloud $\hat{P}\hat{C}$ using the estimated extrinsic matrix:

$$[\hat{X}, \hat{Y}, \hat{Z}, 1]^T = T[X, Y, Z, 1]^T. \quad (21)$$

Most deep learning works exploit the Lie algebra to parameterize the calibration camera-LiDAR extrinsic parameters. In particular, the output of the calibration network is a 1×6 vector $\xi = (v, \omega) \in se(3)$ in which v is the translation vector, and ω is the rotation vector. To recover the original objectives, the rotation vector in $so(3)$ should be transformed to its corresponding rotation matrix. Supposed that $\omega = (\omega_1, \omega_2, \omega_3)^T$, an element $\omega \in so(3)$ can be transformed to $SO(3)$ using the exponential map by:

$$\exp : so(3) \rightarrow SO(3); \hat{\omega} \mapsto e^{\hat{\omega}}, \quad (22)$$

where $\hat{\omega}$ and $e^{\hat{\omega}}$ denote the skew-symmetric matrix from ω and Taylor series expansion for the matrix exponential function, respectively. Then, the rotation matrix can be formed in $SO(3)$, and its Rodrigues formula is derived from the above equation by:

$$R = e^{\hat{\omega}} = I + \frac{\hat{\omega}}{\|\hat{\omega}\|} \sin \|\hat{\omega}\| + \frac{\hat{\omega}^2}{\|\hat{\omega}\|^2} (1 - \cos(\|\hat{\omega}\|)). \quad (23)$$

Thus, the 3D rigid body transformation $T \in SE(3)$ between camera and LiDAR can be represented by:

$$T = \begin{pmatrix} R & t \\ 0 & 1 \end{pmatrix} \text{ where } R \in SO(3), t \triangleq v \in \mathbb{R}^3. \quad (24)$$

3 BENCHMARK EVALUATION

To intuitively demonstrate the effectiveness of the existing learning-based camera calibration methods, we provide the quantitative and qualitative evaluations on representative methods using our benchmark. Considering the application gap and comparison integrity, we also collect and summarize the reported metrics from previous works of certain camera models.

As listed in Tab. 1, the quantitative evaluations on standard model calibration methods are provided. They are summarized from previous methods [23] [113] [7] and evaluated with the Google Street View dataset. The angular errors of predicted camera Up direction, pitch, roll, and FoV are calculated in terms of the mean error (*Mean*) and median error (*Med*). AUC denotes the area under the curve of the cumulative distribution of the horizon line error.

Tab. 2 and Fig. 4 show the quantitative evaluation and qualitative evaluation of representative works for calibrating the distortion camera model, especially in radial distortion. In particular, we split the benchmark into two test sets (Test1 and Test2) based on the square and circle of the wide-angle image. Two types of metrics, namely, image-level metrics (PSNR, SSIM, MS-SSIM) and perception-level metrics (FID, LPIPS-Alex, LPIPS-Vgg) are used to calculate the errors between the corrected wide-angle image and the ground truth. Since FishFormer [68] is customized for the fisheye image with a circular shape, we omit its quantitative metric on Test1 and use the original input as its placeholder in Fig. 4.

We evaluate the performance of representative methods for the cross-view camera model in Tab. 3. The mean average corner error (MACE) is provided for each method. For IHN [87], we test it from two scales following the same setting reported in the paper. Moreover, the point matching errors (PME) of representative methods are listed in Tab. 4, which are collected and summarized from previous works [115], [123] and evaluated on the GHOF [123] test set. Five categories are split to exhibit the performance of different scenes, including regular (RE), foggy (FOG), low light (LL), rainy (RAIN), and snowy (SNOW) scenes.

We also evaluate the performance of several open source methods on cross-sensor camera model in Tab. 5. Experimental results are reported according to the metrics of rotation and translation. For the rotation part, we calculate quaternion angle distance and absolute angle error of Euler angles including Roll, Pitch, and Yaw. For the translation part, we compute the Euclidean distance between the predicted translation vector and ground truth with absolute error in X , Y , Z directions. Since there has not been a common benchmark for camera-LiDAR calibration yet, all selected methods were trained on the KITTI [117] dataset with different specific settings. In the evaluation, we devote ourselves to aligning the proposed benchmark to each method's original setting. Due to the difference in image size, image content, LiDAR beam number, and procedure of generating input depth map, the input color image and depth map to different trained models are inconsistent. From Tab. 5, Fig. 5, and Fig. 6, we can observe there are noticeable performance degradations of comparison methods under the cross-domain evaluations.

4 MORE FUTURE DIRECTIONS

4.1 Dataset

One of the main challenges of learning-based camera calibrations is the difficulty in constructing datasets with high accuracy. This requires laborious manual intervention to obtain real-world data with labels. As we summarized, approximately 70% of the works rely on synthesized datasets. However, the significant differences between synthesized and real-world datasets cannot be ignored, leading to domain gaps in the learned models. Therefore, the construction of a standardized, large-scale calibration dataset would significantly benefit this community. Recent works have demonstrated that well-designed learning strategies, such as semi-supervised learning [62], self-supervised learning [17], [77], and unsupervised learning [74], [81], can help address the demand for annotations in learning-based camera calibrations. These strategies also have the potential to discover additional calibration priors within the data itself.

4.2 Transfer learning

The advancements in deep learning have led to the development of transfer learning techniques, which could facilitate the transfer of knowledge learned from one camera to another. This approach can significantly speed up and streamline the calibration process, making it more efficient and cost-effective. Transfer learning can be especially useful in applications that involve multiple cameras or mobile

TABLE 1
Quantitative evaluation on the representative camera calibration methods for the standard model.

Method	Up Direction ($^{\circ}$) \downarrow		Pitch ($^{\circ}$) \downarrow		Roll ($^{\circ}$) \downarrow		FoV ($^{\circ}$) \downarrow		AUC (%) \uparrow
	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.	
DeepHorizon [3]	3.58	3.01	2.76	2.12	1.78	1.67	-	-	80.29
Hold-Geoffroy et al. [28]	2.73	2.13	2.39	1.78	0.96	0.66	4.61	3.89	80.40
UprightNet [6]	28.20	26.10	26.56	24.56	6.22	4.33	-	-	-
Lee et al. [7]	2.12	1.61	1.92	1.38	0.75	0.47	6.01	3.72	83.12
CTRL-C [23]	1.80	1.52	1.58	1.31	0.66	0.53	3.59	2.72	87.29
PerspectiveNet [113]	-	-	1.36	1.18	0.66	0.52	3.07	2.33	-

TABLE 2
Qualitative evaluation on the representative camera calibration methods for distortion model.

Comparison on Test1		Metrics					
Methods	Type	PSNR \uparrow	SSIM \uparrow	MS-SSIM \uparrow	FID \downarrow	LPIPS-Alex \downarrow	LPIPS-Vgg \downarrow
DCCNN [27]	Regression	9.18	0.1583	0.1617	360.90	0.4081	0.5258
DeepCalib [29]	Regression	10.71	0.2289	0.2534	161.31	0.3330	0.4482
FishFormer [68]	Regression	-	-	-	-	-	-
BlindCor [36]	Regression	8.02	0.1356	0.1392	400.77	0.4956	0.5536
DR-GAN [33]	Generation	16.81	0.5058	0.6794	148.50	0.1782	0.3268
DDM [42]	Generation	17.43	0.5659	0.7191	125.34	0.1455	0.2333
PCN [56]	Generation	24.59	0.8726	0.9594	42.78	0.0388	0.0671
Comparison on Test2		Metrics					
Methods	Type	PSNR \uparrow	SSIM \uparrow	MS-SSIM \uparrow	FID \downarrow	LPIPS-Alex \downarrow	LPIPS-Vgg \downarrow
DCCNN [27]	Regression	15.23	0.4205	0.3609	125.99	0.1586	0.2110
DeepCalib [29]	Regression	11.47	0.2715	0.3655	139.26	0.2853	0.3886
FishFormer [68]	Regression	21.52	0.7540	0.8971	66.37	0.1133	0.1107
BlindCor [36]	Generation	12.01	0.3075	0.2531	133.16	0.3595	0.3907
DR-GAN [33]	Generation	17.31	0.5133	0.6901	116.78	0.1218	0.2056
DDM [42]	Generation	18.84	0.6247	0.7614	78.38	0.1055	0.1894
PCN [56]	Generation	21.28	0.7027	0.8595	54.64	0.0812	0.1088

TABLE 3
Qualitative evaluation on the representative methods for cross-view camera model.

Method	MSCOCO [114]	GoogleEarth [82]	GoogleMap [82]	CA-Homo [81]
MHN [80]	1.1512	10.3078	13.1610	-
MHN+DLKFM [82]	0.7687	3.9629	5.2664	-
IHN-scale1 [87]	0.2652	1.5135	0.9610	-
IHN-scale2 [87]	0.1234	1.2110	0.6751	-
CA-UDHN [81]	-	-	-	0.8605
BasesHomo [84]	-	-	-	0.6808
HomoGAN [88]	-	-	-	0.3651

TABLE 4
The point matching errors (PME) of representative methods for cross-view camera model.

Methods	AVG	RE	FOG	LL	RAIN	SNOW
$\mathcal{I}_{3 \times 3}$	6.33	4.94	7.24	8.09	5.48	5.89
CA-UDHN [81]	3.87	4.10	3.84	6.99	1.27	3.17
BasesHomo [84]	2.28	2.02	1.43	4.90	0.78	2.29
HomoGAN [88]	1.95	1.73	0.60	3.95	0.47	3.02
DHN [69]	6.61	6.04	6.02	7.68	6.99	6.32
LocalTrans [83]	5.72	4.06	6.49	5.95	5.78	6.34
IHN [87]	8.17	7.10	8.71	9.34	6.57	9.13
DHN* [69]	3.01	1.92	3.94	4.54	1.98	2.66
LocalTrans* [83]	2.89	1.78	4.27	4.59	1.37	2.43
IHN* [87]	2.59	2.21	3.05	4.70	0.98	2.03
RealSH [115]	1.72	1.60	0.88	4.42	0.43	1.28

devices. For example, in a multi-camera system, transfer learning can be used to calibrate all the cameras using the data collected from a single camera, reducing the time and effort required for calibration. Similarly, in mobile devices, transfer learning can enable faster and more accurate calibration of the camera, resulting in improved image quality and performance.

4.3 Robustness to noise and outliers

Another promising application of deep learning in camera calibration is improving the robustness of calibration to noise and outliers in the data. This approach can help ensure accurate calibration even in challenging environments, with low-quality data or noisy sensor readings. Conventionally, camera calibration algorithms are sensitive to noise and outliers in the data, which can lead to significant errors in the

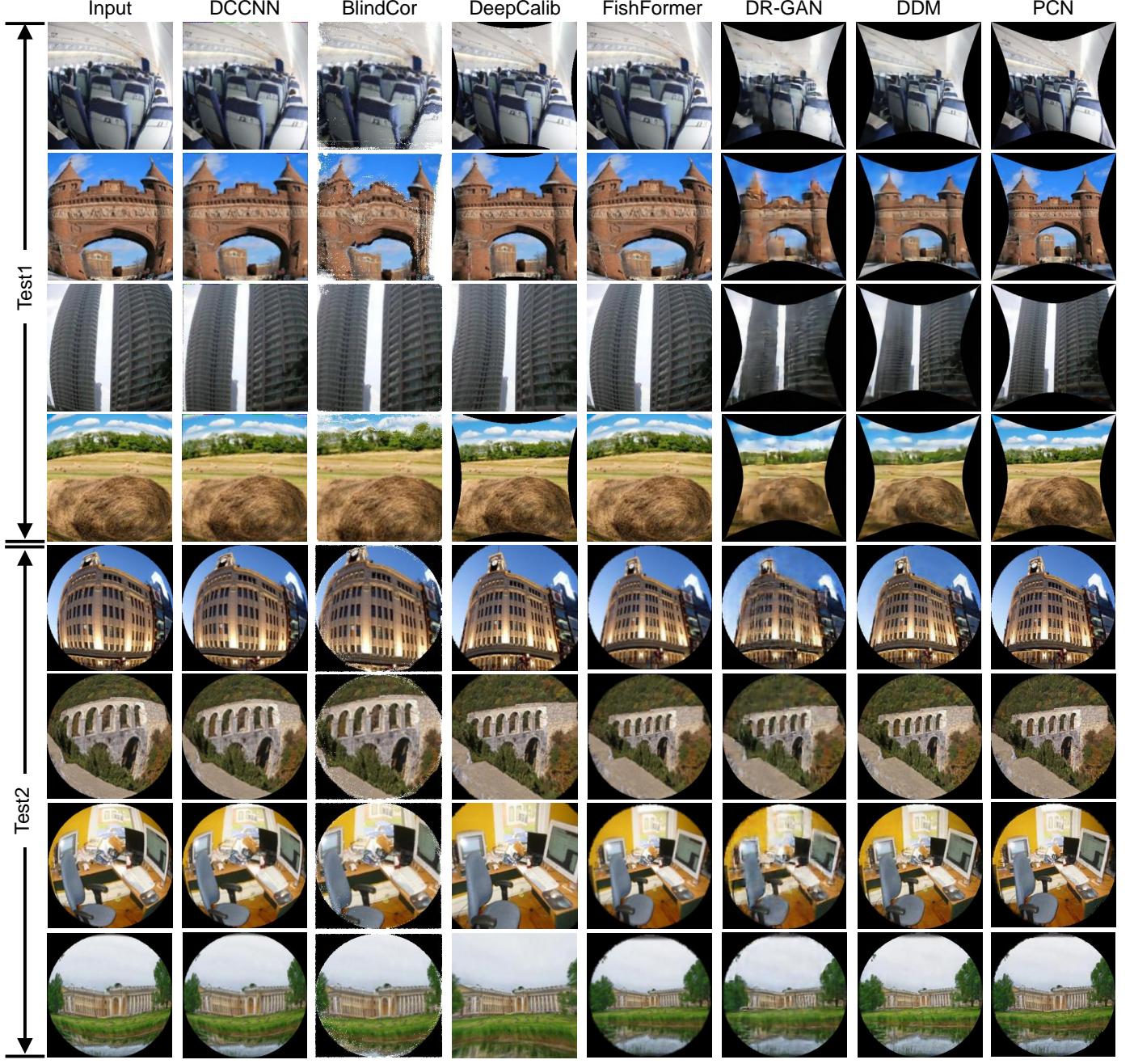


Fig. 4. Qualitative evaluation on the representative camera calibration methods for distortion model.

estimated camera parameters. However, with the application of deep learning, it is possible to learn more robust and accurate models that can better handle noise and outliers in the data. For instance, regularization techniques can be used to impose constraints on the learned parameters, preventing overfitting and enhancing the generalization ability of the model. Moreover, outlier detection techniques can be used to identify and exclude data points that are likely to be outliers, reducing their impact on the calibration process. This can be achieved using various statistical and machine-learning methods, such as clustering, classification, and regression.

4.4 Online calibration

With the rapid development of deep learning, online camera calibration is becoming more efficient and practical. This technique involves updating the calibration parameters in real-time, allowing for better performance as the camera moves or as the environment changes. This can be achieved using deep learning algorithms that can learn the complex relationships between the camera parameters and the image data. Learning-based camera calibration has the potential to revolutionize various industries, such as robotics and augmented reality. In robotics, online calibration can improve the accuracy of robot vision, which is crucial for tasks such as object detection and manipulation. Similarly, in augmented reality, online calibration can enhance the

TABLE 5
Qualitative evaluation on the representative calibration methods for Cross-Sensor model.

Methods		E_t	Translation (cm)			E_R	Rotation (°)		
			X	Y	Z		Roll	Pitch	Yaw
CMRNet [116] 10° / 2.0 m	KITTI [117]	126.2820	72.8394	55.3118	77.1015	12.9863	5.7665	5.3867	3.9606
	Apollo [118], [119]	136.7067	44.6239	109.4236	82.5008	12.2843	8.2973	6.6979	5.7054
	NuScenes [120]	124.7204	43.6849	69.5890	64.1339	11.4339	9.9125	9.3100	5.3005
	ONCE [121]	145.6419	83.0297	26.1744	70.9436	11.3011	7.2812	7.4463	3.2014
	DAIR-V2X [122]	158.3873	51.8842	85.4628	79.1833	10.3119	3.1635	3.7430	5.5995
CalibNet [91] 10° / 0.2 m	KITTI [117]	72.3976	37.4153	46.3160	30.4167	10.5577	4.8652	7.3470	5.4789
	Apollo [118], [119]	71.6006	18.7960	50.2996	16.5990	11.0372	6.0806	5.4223	9.4484
	NuScenes [120]	73.4821	52.6383	45.7675	27.4490	11.5573	8.0735	8.4622	5.6564
	ONCE [121]	74.0186	46.3395	49.3340	31.7829	12.4912	5.3240	5.3686	6.7609
	DAIR-V2X [122]	77.3034	41.9963	54.7920	51.8221	11.6145	7.4061	1.7497	6.4258
LCCNet [97] 20° / 1.5 m	KITTI [117]	45.0945	9.2671	7.2877	11.6077	1.8611	0.6823	0.5427	0.2704
	Apollo [118], [119]	159.9080	45.5809	67.7147	106.1134	15.1098	2.5412	3.8089	2.8912
	NuScenes [120]	232.1684	73.4198	120.1943	98.3628	12.3146	2.5942	3.7928	9.0684
	ONCE [121]	185.0779	74.6479	72.9388	74.6713	11.5796	8.6251	6.2554	4.7137
	DAIR-V2X [122]	228.1306	113.5558	86.6683	98.3726	7.3732	2.0581	3.7936	4.7334
RGNet [92] 20° / 0.3 m	KITTI [117]	77.7839	12.5584	14.4128	68.0927	9.8177	2.1106	2.6808	6.7095
	Apollo [118], [119]	542.7074	215.6324	254.2824	218.8926	35.4403	17.8385	17.7508	18.2130
	NuScenes [120]	92.3324	44.9838	51.1874	51.5143	35.5249	22.6364	25.5031	18.2290
	ONCE [121]	73.8569	55.6384	49.2977	33.9010	26.0675	6.7498	6.7093	20.2362
	DAIR-V2X [122]	397.6241	260.4813	208.1069	187.8416	34.8505	22.0118	23.7622	18.6647

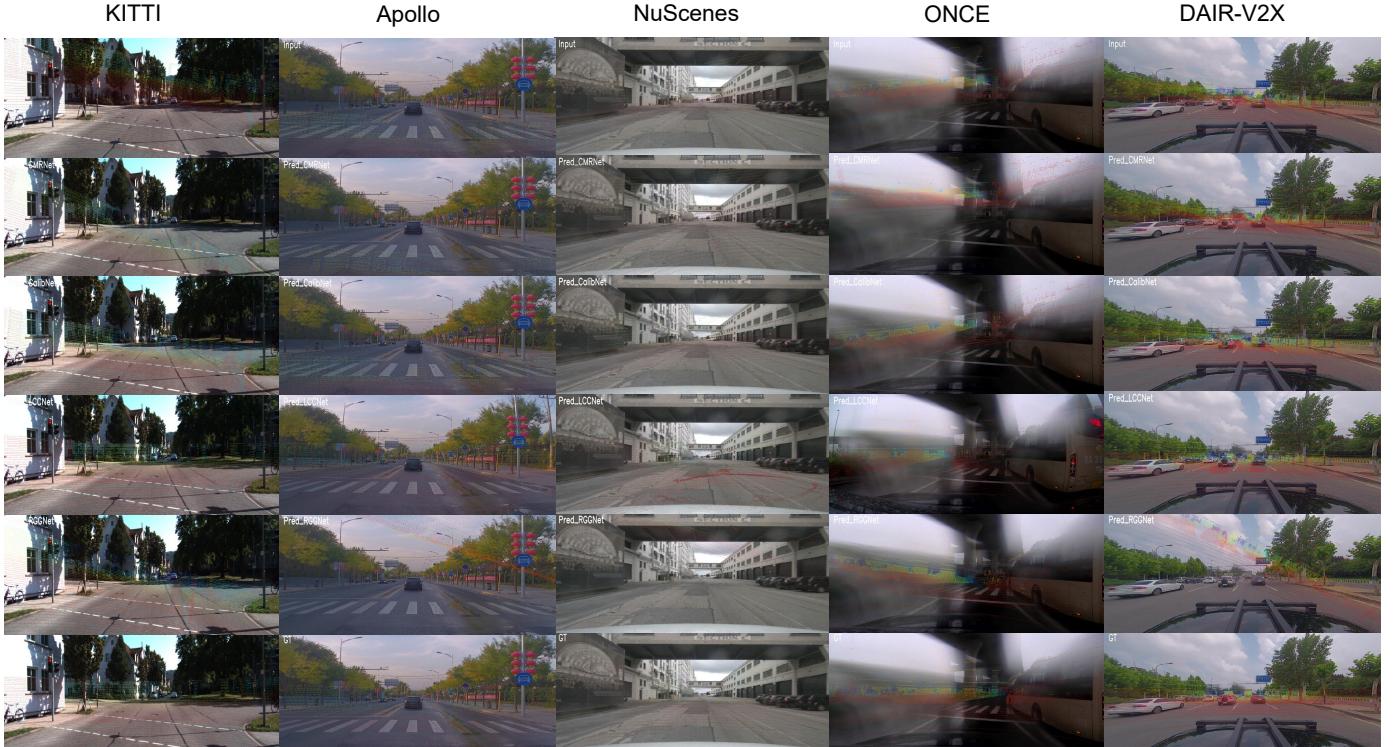


Fig. 5. Qualitative evaluation on the representative calibration methods for cross-sensor model.

user experience by ensuring that virtual objects are correctly aligned with the real world. This can help create more realistic and immersive AR applications, which have numerous practical applications in fields such as entertainment, education, and training.

4.5 Multimodal calibration

The potential of deep learning techniques in camera calibration goes beyond traditional photography and computer

vision applications. It could also be applied to calibrate cameras with other sensors, such as remote sensing, infrared sensors, or radar. This advancement could lead to more precise and robust perception in various applications, including but not limited to autonomous driving, where multiple sensors are used. Incorporating deep learning-based calibration methods with multiple sensors could enhance the accuracy of the fusion of data from different sources. It could facilitate more accurate perception in challenging

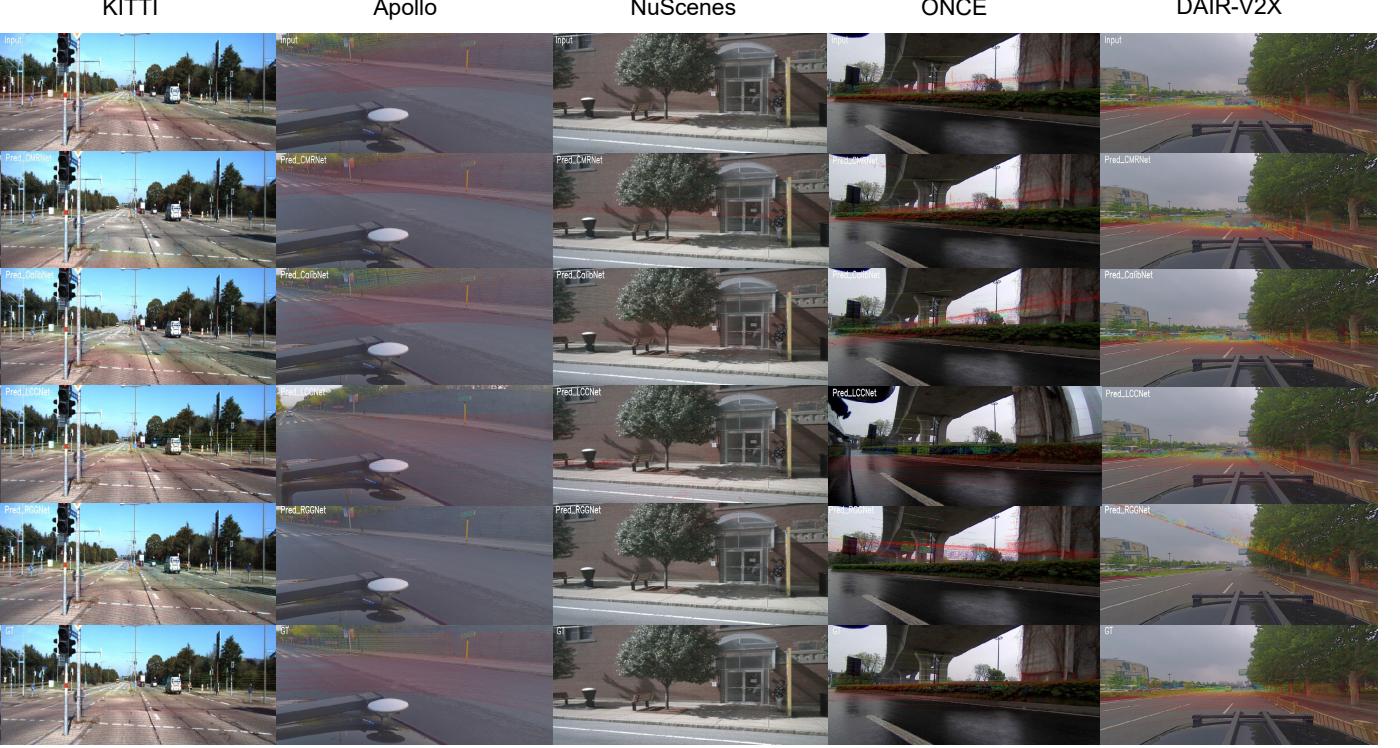


Fig. 6. Qualitative evaluation on the representative calibration methods for cross-sensor model.

environments such as low-light conditions, occlusions, and adverse weather conditions. Furthermore, the ability to calibrate multiple sensors with deep learning methods could provide more reliable and consistent results compared to traditional calibration techniques.

These are a few potential directions for future research in camera calibration with deep learning. As the field continues to evolve, there may be many other exciting avenues for exploration and innovation. In addition, it is also thrilling to see how this technology will continue to impact various industries in the future.

REFERENCES

- [1] S. Workman, C. Greenwell, M. Zhai, R. Baltenberger, and N. Jacobs, "Deepfocal: A method for direct focal length estimation," in *IEEE International Conference on Image Processing*, 2015, pp. 1369–1373.
- [2] A. Kendall, M. Grimes, and R. Cipolla, "Posenet: A convolutional network for real-time 6-dof camera relocalization," in *International Conference on Computer Vision*, 2015.
- [3] S. Workman, M. Zhai, and N. Jacobs, "Horizon lines in the wild," *arXiv preprint arXiv:1604.02129*, 2016.
- [4] M. Zhai, S. Workman, and N. Jacobs, "Detecting vanishing points using global image context in a non-manhattan world," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.
- [5] C.-K. Chang, J. Zhao, and L. Itti, "Deepvp: Deep learning for vanishing point detection on 1 million street view images," in *IEEE International Conference on Robotics and Automation*, 2018, pp. 4496–4503.
- [6] W. Xian, Z. Li, M. Fisher, J. Eisenmann, E. Shechtman, and N. Snavely, "Uprightnet: Geometry-aware camera orientation estimation from single images," in *International Conference on Computer Vision*, 2019.
- [7] J. Lee, M. Sung, H. Lee, and J. Kim, "Neural geometric parser for single image camera calibration," in *European Conference on Computer Vision*, 2020, pp. 541–557.
- [8] Y. Zhou, H. Qi, J. Huang, and Y. Ma, "Neurvps: Neural vanishing point scanning via conic convolution," *Advances in Neural Information Processing Systems*, vol. 32, 2019.
- [9] R. Jung, A. S. J. Lee, A. Ashtari, and J.-C. Bazin, "Deep360up: A deep learning-based approach for automatic vr image upright adjustment," in *IEEE Conference on Virtual Reality and 3D User Interfaces*, 2019.
- [10] B. Davidson, M. S. Alvi, and J. F. Henriques, "360° camera alignment via segmentation," in *European Conference on Computer Vision*, 2020, pp. 579–595.
- [11] Y.-Y. Jau, R. Zhu, H. Su, and M. Chandraker, "Deep keypoint-based camera pose estimation with geometric constraints," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 4950–4957.
- [12] M. Baradad and A. Torralba, "Height and uprightness invariance for 3d prediction from a single view," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.
- [13] Q. Zheng, J. Chen, Z. Lu, B. Shi, X. Jiang, K.-H. Yap, L.-Y. Duan, and A. C. Kot, "What does plate glass reveal about camera calibration?" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.
- [14] R. Zhu, X. Yang, Y. Hold-Geoffroy, F. Perazzi, J. Eisenmann, K. Sunkavalli, and M. Chandraker, "Single view metrology in the wild," in *European Conference on Computer Vision*, 2020, pp. 316–333.
- [15] Y. Gil, S. Elmaleem, H. Haim, E. Marom, and R. Giryes, "Online training of stereo self-calibration using monocular depth estimation," *IEEE Transactions on Computational Imaging*, vol. 7, pp. 812–823, 2021.
- [16] S. Garg, D. P. Mohanty, S. P. Thota, and S. Moharana, "A simple approach to image tilt correction with self-attention mobilenet for smartphones," *arXiv preprint arXiv:2111.00398*, 2021.
- [17] J. Fang, I. Vasiljevic, V. Guizilini, R. Ambrus, G. Shakhnarovich, A. Gaidon, and M. R. Walter, "Self-supervised camera self-calibration from video," *arXiv preprint arXiv:2112.03325*, 2021.
- [18] T. H. Butt and M. Taj, "Camera calibration through camera projection loss," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022, pp. 2649–2653.
- [19] G. Ponimatkin, Y. Labb  , B. Russell, M. Aubry, and J. Sivic, "Focal length and object pose estimation via render and compare," in

- Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3825–3834.
- [20] K. Chen, N. Snavely, and A. Makadia, “Wide-baseline relative camera pose estimation with directional learning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3258–3268.
- [21] Y. Lin, R. Wiersma, S. L. Pintea, K. Hildebrandt, E. Eisemann, and J. C. van Gemert, “Deep vanishing point detection: Geometric priors make dataset variations vanish,” *arXiv preprint arXiv:2203.08586*, 2022.
- [22] T. Do, O. Miksik, J. DeGol, H. S. Park, and S. N. Sinha, “Learning to detect scene landmarks for camera localization,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 132–11 142.
- [23] J. Lee, H. Go, H. Lee, S. Cho, M. Sung, and J. Kim, “Ctrl-c: Camera calibration transformer with line-classification,” in *International Conference on Computer Vision*, 2021, pp. 16 228–16 237.
- [24] M. Kocabas, C.-H. P. Huang, J. Tesch, L. Müller, O. Hilliges, and M. J. Black, “Spec: Seeing people in the wild with an estimated camera,” in *International Conference on Computer Vision*, 2021, pp. 11 035–11 045.
- [25] C. M. Parameshwara, G. Hari, C. Fermüller, N. J. Sanket, and Y. Aloimonos, “Diffposenet: Direct differentiable camera pose estimation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 6845–6854.
- [26] L. Yang, R. Shrestha, W. Li, S. Liu, G. Zhang, Z. Cui, and P. Tan, “Scenesqueezee: Learning to compress scene for camera relocation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 8259–8268.
- [27] J. Rong, S. Huang, Z. Shang, and X. Ying, “Radial lens distortion correction using convolutional neural networks trained with synthesized images,” in *Asian Conference on Computer Vision*, 2016, pp. 35–49.
- [28] Y. Hold-Geoffroy, K. Sunkavalli, J. Eisenmann, M. Fisher, E. Gambaretto, S. Hadap, and J.-F. Lalonde, “A perceptual measure for deep single image camera calibration,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.
- [29] O. Bogdan, V. Eckstein, F. Rameau, and J.-C. Bazin, “Deepcalib: a deep learning approach for automatic intrinsic calibration of wide field-of-view cameras,” in *Proceedings of the 15th ACM SIGGRAPH European Conference on Visual Media Production*, 2018.
- [30] V. Rengarajan, Y. Balaji, and A. Rajagopalan, “Unrolling the shutter: Cnn to correct motion distortions,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2291–2299.
- [31] X. Yin, X. Wang, J. Yu, M. Zhang, P. Fua, and D. Tao, “Fisheye-erectnet: A multi-context collaborative deep network for fisheye image rectification,” in *European Conference on Computer Vision*, 2018.
- [32] Y. Shi, D. Zhang, J. Wen, X. Tong, X. Ying, and H. Zha, “Radial lens distortion correction by adding a weight layer with inverted foveal models to convolutional neural networks,” in *International Conference on Pattern Recognition*, 2018.
- [33] K. Liao, C. Lin, Y. Zhao, and M. Gabbouj, “Dr-gan: Automatic radial distortion rectification using conditional gan in real-time,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 3, pp. 725–733, 2020.
- [34] ——, “Distortion rectification from static to dynamic: A distortion sequence construction perspective,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 11, pp. 3870–3882, 2020.
- [35] S. Yang, C. Lin, K. Liao, Y. Zhao, and M. Liu, “Unsupervised fisheye image correction through bidirectional loss with geometric prior,” *Journal of Visual Communication and Image Representation*, vol. 66, p. 102692, 2020.
- [36] X. Li, B. Zhang, P. V. Sander, and J. Liao, “Blind geometric distortion correction on images through deep learning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.
- [37] B. Zhuang, Q.-H. Tran, P. Ji, L.-F. Cheong, and M. Chandraker, “Learning structure-and-motion-aware rolling shutter correction,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.
- [38] Z. Xue, N. Xue, G.-S. Xia, and W. Shen, “Learning to calibrate straight lines for fisheye image rectification,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.
- [39] B. Zhuang, Q.-H. Tran, G. H. Lee, L. F. Cheong, and M. Chandraker, “Degeneracy in self-calibration revisited and a deep learning solution for uncalibrated slam,” in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2019, pp. 3766–3773.
- [40] M. Lopez, R. Mari, P. Gargallo, Y. Kuang, J. Gonzalez-Jimenez, and G. Haro, “Deep single image camera calibration with radial distortion,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.
- [41] Y. Zhao, Z. Huang, T. Li, W. Chen, C. LeGendre, X. Ren, A. Shapiro, and H. Li, “Learning perspective undistortion of portraits,” in *International Conference on Computer Vision*, 2019.
- [42] K. Liao, C. Lin, Y. Zhao, and M. Xu, “Model-free distortion rectification framework bridged by distortion distribution map,” *IEEE Transactions on Image Processing*, vol. 29, pp. 3707–3718, 2020.
- [43] A. Cramariuc, A. Petrov, R. Suri, M. Mittal, R. Siegwart, and C. Cadena, “Learning camera miscalibration detection,” in *IEEE International Conference on Robotics and Automation*, 2020, pp. 4997–5003.
- [44] C. Zhang, F. Rameau, J. Kim, D. M. Argaw, J.-C. Bazin, and I. S. Kweon, “Deepptz: Deep self-calibration for ptz cameras,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020.
- [45] C.-H. Chao, P.-L. Hsu, H.-Y. Lee, and Y.-C. F. Wang, “Self-supervised deep learning for fisheye image rectification,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020, pp. 2248–2252.
- [46] Y. Shi, X. Tong, J. Wen, H. Zhao, X. Ying, and H. Zha, “Position-aware and symmetry enhanced gan for radial distortion correction,” in *International Conference on Pattern Recognition*, 2021, pp. 1701–1708.
- [47] H. Zhao, Y. Shi, X. Tong, X. Ying, and H. Zha, “A simple yet effective pipeline for radial distortion correction,” in *IEEE International Conference on Image Processing*, 2020, pp. 878–882.
- [48] Y.-H. Li, I.-C. Lo, and H. H. Chen, “Deep face rectification for 360° dual-fisheye cameras,” *IEEE Transactions on Image Processing*, vol. 30, pp. 264–276, 2021.
- [49] H. Zhao, X. Ying, Y. Shi, X. Tong, J. Wen, and H. Zha, “Rdcface: Radial distortion correction for face recognition,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.
- [50] Z.-C. Xue, N. Xue, and G.-S. Xia, “Fisheye distortion rectification from deep straight lines,” *arXiv preprint arXiv:2003.11386*, 2020.
- [51] P. Liu, Z. Cui, V. Larsson, and M. Pollefeys, “Deep shutter unrolling network,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5941–5949.
- [52] K. Liao, C. Lin, and Y. Zhao, “A deep ordinal distortion estimation approach for distortion rectification,” *IEEE Transactions on Image Processing*, vol. 30, pp. 3362–3375, 2021.
- [53] K. Zhao, C. Lin, K. Liao, S. Yang, and Y. Zhao, “Revisiting radial distortion rectification in polar-coordinates: A new and efficient learning perspective,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 6, pp. 3552–3560, 2021.
- [54] J. Zhao, S. Wei, L. Liao, and Y. Zhao, “Dqn-based gradual fisheye image rectification,” *Pattern Recognition Letters*, vol. 152, pp. 129–134, 2021.
- [55] J. Tan, S. Zhao, P. Xiong, J. Liu, H. Fan, and S. Liu, “Practical wide-angle portraits correction with deep structured models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3498–3506.
- [56] S. Yang, C. Lin, K. Liao, C. Zhang, and Y. Zhao, “Progressively complementary network for fisheye image rectification using appearance flow,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6348–6357.
- [57] J. Fan, J. Zhang, and D. Tao, “Sir: Self-supervised image rectification via seeing the same scene from multiple different lenses,” *IEEE Transactions on Image Processing*, 2022.
- [58] K. Liao, C. Lin, L. Liao, Y. Zhao, and W. Lin, “Multi-level curriculum for training a distortion-aware barrel distortion rectification model,” in *International Conference on Computer Vision*, 2021, pp. 4389–4398.
- [59] N. Wakai and T. Yamashita, “Deep single fisheye image camera calibration for over 180-degree projection of field of view,” in *International Conference on Computer Vision Workshops*, 2021, pp. 1174–1183.
- [60] N. Wakai, S. Sato, Y. Ishii, and T. Yamashita, “Rethinking generic camera models for deep single image camera calibration to

- recover rotation and fisheye distortion," in *European Conference on Computer Vision*, vol. 13678, 2022, pp. 679–698.
- [61] Z. Zhong, Y. Zheng, and I. Sato, "Towards rolling shutter correction and deblurring in dynamic scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9219–9228.
- [62] F. Zhu, S. Zhao, P. Wang, H. Wang, H. Yan, and S. Liu, "Semi-supervised wide-angle portraits correction by multi-scale transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 19 689–19 698.
- [63] M. Cao, Z. Zhong, J. Wang, Y. Zheng, and Y. Yang, "Learning adaptive warping for real-world rolling shutter correction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17 785–17 793.
- [64] X. Zhou, P. Duan, Y. Ma, and B. Shi, "Evnunroll: Neuromorphic events based rolling shutter image correction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17 775–17 784.
- [65] B. Fan and Y. Dai, "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*, 2021, pp. 4228–4237.
- [66] B. Fan, Y. Dai, and M. He, "Sunet: symmetric undistortion network for rolling shutter correction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 4541–4550.
- [67] Y. Zhang, X. Zhao, and D. Qian, "Learning-based framework for camera calibration with distortion correction and high precision feature detection," *arXiv preprint arXiv:2202.00158*, 2022.
- [68] Y. Shangrong, L. Chunyu, L. Kang, and Z. Yao, "Fishformer: Annulus slicing-based transformer for fisheye rectification with efficacy domain exploration," *arXiv preprint arXiv:2207.01925*, 2022.
- [69] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Deep image homography estimation," *arXiv preprint arXiv:1606.03798*, 2016.
- [70] C.-H. Chang, C.-N. Chou, and E. Y. Chang, "Clnk: Cascaded lucas-kanade networks for image alignment," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.
- [71] F. Erlik Nowruzi, R. Laganiere, and N. Japkowicz, "Homography estimation from image pairs with hierarchical convolutional networks," in *International Conference on Computer Vision Workshops*, 2017.
- [72] R. Ranftl and V. Koltun, "Deep fundamental matrix estimation," in *European Conference on Computer Vision*, 2018.
- [73] O. Poursaeed, G. Yang, A. Prakash, Q. Fang, H. Jiang, B. Hariharan, and S. Belongie, "Deep fundamental matrix estimation without correspondences," in *European Conference on Computer Vision Workshops*, 2018.
- [74] T. Nguyen, S. W. Chen, S. S. Shivakumar, C. J. Taylor, and V. Kumar, "Unsupervised deep homography: A fast and robust homography estimation model," *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 2346–2353, 2018.
- [75] R. Zeng, S. Denman, S. Sridharan, and C. Fookes, "Rethinking planar homography estimation using perspective fields," in *Asian Conference on Computer Vision*, 2018, pp. 571–586.
- [76] Y. Li, W. Pei, and Z. He, "Srhen: stepwise-refining homography estimation network via parsing geometric correspondences in deep latent space," in *Proceedings of the ACM International Conference on Multimedia*, 2020, pp. 3063–3071.
- [77] X. Wang, C. Wang, B. Liu, X. Zhou, L. Zhang, J. Zheng, and X. Bai, "Multi-view stereo in the deep learning era: A comprehensive review," *Displays*, vol. 70, p. 102102, 2021.
- [78] S. Ammar Abbas and A. Zisserman, "A geometric approach to obtain a bird's eye view from an image," in *International Conference on Computer Vision Workshops*, 2019.
- [79] L. Sha, J. Hobbs, P. Felsen, X. Wei, P. Lucey, and S. Ganguly, "End-to-end camera calibration for broadcast videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.
- [80] H. Le, F. Liu, S. Zhang, and A. Agarwala, "Deep homography estimation for dynamic scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.
- [81] J. Zhang, C. Wang, S. Liu, L. Jia, N. Ye, J. Wang, J. Zhou, and J. Sun, "Content-aware unsupervised deep homography estimation," in *European Conference on Computer Vision*, 2020, pp. 653–669.
- [82] Y. Zhao, X. Huang, and Z. Zhang, "Deep lucas-kanade homography for multimodal image alignment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15 950–15 959.
- [83] R. Shao, G. Wu, Y. Zhou, Y. Fu, L. Fang, and Y. Liu, "Localtrans: A multiscale local transformer network for cross-resolution homography estimation," in *International Conference on Computer Vision*, 2021, pp. 14 890–14 899.
- [84] N. Ye, C. Wang, H. Fan, and S. Liu, "Motion basis learning for unsupervised deep homography estimation with subspace projection," in *International Conference on Computer Vision*, 2021, pp. 13 117–13 125.
- [85] Y. Chen, G. Wang, P. An, Z. You, and X. Huang, "Fast and accurate homography estimation using extendable compression network," in *International Conference on Image Processing*, 2021, pp. 1024–1028.
- [86] L. Nie, C. Lin, K. Liao, S. Liu, and Y. Zhao, "Depth-aware multi-grid deep homography estimation with contextual correlation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 7, pp. 4460–4472, 2021.
- [87] S.-Y. Cao, J. Hu, Z. Sheng, and H.-L. Shen, "Iterative deep homography estimation," *arXiv preprint arXiv:2203.15982*, 2022.
- [88] M. Hong, Y. Lu, N. Ye, C. Lin, Q. Zhao, and S. Liu, "Unsupervised homography estimation with coplanarity-aware gan," *arXiv preprint arXiv:2205.03821*, 2022.
- [89] S. Liu, N. Ye, C. Wang, J. Zhang, L. Jia, K. Luo, J. Wang, and J. Sun, "Content-aware unsupervised deep homography estimation and its extensions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 3, pp. 2849–2863, 2022.
- [90] N. Schneider, F. Piewak, C. Stiller, and U. Franke, "Regnet: Multimodal sensor registration using deep neural networks," in *IEEE intelligent vehicles symposium*, 2017, pp. 1803–1810.
- [91] G. Iyer, R. K. Ram, J. K. Murthy, and K. M. Krishna, "Calibnet: Geometrically supervised extrinsic calibration using 3d spatial transformer networks," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2018, pp. 1110–1117.
- [92] K. Yuan, Z. Guo, and Z. J. Wang, "Rggnet: Tolerance aware lidar-camera online calibration with geometric deep learning and generative model," *IEEE Robotics and Automation Letters*, vol. 5, no. 4, pp. 6956–6963, 2020.
- [93] Y. Zhu, C. Li, and Y. Zhang, "Online camera-lidar calibration with sensor semantic information," in *IEEE International Conference on Robotics and Automation*, 2020, pp. 4970–4976.
- [94] W. Wang, S. Nobuhara, R. Nakamura, and K. Sakurada, "Soic: Semantic online initialization and calibration for lidar and camera," *arXiv preprint arXiv:2003.04260*, 2020.
- [95] J. Shi, Z. Zhu, J. Zhang, R. Liu, Z. Wang, S. Chen, and H. Liu, "Calibrcnn: Calibrating camera and lidar by recurrent convolutional neural network and geometric constraints," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2020, pp. 10 197–10 202.
- [96] S. Wu, A. Hadachi, D. Vivet, and Y. Prabhakar, "Netcalib: A novel approach for lidar-camera auto-calibration based on deep learning," in *International Conference on Pattern Recognition*, 2021, pp. 6648–6655.
- [97] X. Lv, B. Wang, Z. Dou, D. Ye, and S. Wang, "Lccnet: Lidar and camera self-calibration using cost volume network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2894–2901.
- [98] X. Lv, S. Wang, and D. Ye, "Cfnet: Lidar-camera registration using calibration flow network," *Sensors*, vol. 21, no. 23, p. 8112, 2021.
- [99] Z. Liu, H. Tang, S. Zhu, and S. Han, "Semalign: Annotation-free camera-lidar calibration with semantic alignment loss," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2021, pp. 8845–8851.
- [100] X. Jing, X. Ding, R. Xiong, H. Deng, and Y. Wang, "Dxq-net: Differentiable lidar-camera extrinsic calibration using quality-aware flow," *arXiv preprint arXiv:2203.09385*, 2022.
- [101] K. Akio, Z. Yiyang, Z. Pengwei, Z. Wei, and T. Masayoshi, "Sst-calib: Simultaneous spatial-temporal parameter calibration between lidar and camera," *arXiv preprint arXiv:2207.03704*, 2022.
- [102] Y. Sun, J. Li, Y. Wang, X. Xu, X. Yang, and Z. Sun, "Atop: An attention-to-optimization approach for automatic lidar-camera calibration via cross-modal object matching," *IEEE Transactions on Intelligent Vehicles*, 2022.
- [103] G. Wang, J. Qiu, Y. Guo, and H. Wang, "Fusionnet: Coarse-to-fine extrinsic calibration network of lidar and camera with hierarchi-

- cal point-pixel fusion," in *International Conference on Robotics and Automation*, 2022, pp. 8964–8970.
- [104] C. Ye, H. Pan, and H. Gao, "Keypoint-based lidar-camera online calibration with robust geometric network," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–11, 2021.
- [105] S. T. Barnard, "Interpreting perspective images," *Artificial Intelligence*, vol. 21, no. 4, pp. 435–462, 1983.
- [106] J. Fan, J. Zhang, S. J. Maybank, and D. Tao, "Wide-angle image rectification: a survey," *International Journal of Computer Vision*, vol. 130, no. 3, pp. 747–776, 2022.
- [107] A. Wang, T. Qiu, and L. Shao, "A simple method of radial distortion correction with centre of distortion estimation," *Journal of Mathematical Imaging and Vision*, vol. 35, no. 3, pp. 165–172, 2009.
- [108] A. W. Fitzgibbon, "Simultaneous linear estimation of multiple view geometry and lens distortion," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2001.
- [109] O. Ait-Aider, N. Andreff, J. M. Lavest, and P. Martinet, "Simultaneous object pose and velocity computation using a single view from a rolling shutter camera," in *European Conference on Computer Vision*, 2006, pp. 56–68.
- [110] H. C. Longuet-Higgins, "A computer algorithm for reconstructing a scene from two projections," *Nature*, vol. 293, no. 5828, pp. 133–135, 1981.
- [111] S. Baker, A. Datta, and T. Kanade, "Parameterizing homographies," *Robotics Institute, Pittsburgh, PA, Tech. Rep. CMU-RI-TR-06-11*, 2006.
- [112] B. K. Horn, "The direct linear transformation from comparator coordinates into object-space coordinates in close-range photogrammetry," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 42, no. 3, pp. 125–133, 1987.
- [113] L. Jin, J. Zhang, Y. Hold-Geoffroy, O. Wang, K. Blackburn-Matzen, M. Stich, and D. F. Fouhey, "Perspective fields for single image camera calibration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 17307–17316.
- [114] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*, 2014, pp. 740–755.
- [115] H. Jiang, H. Li, S. Han, H. Fan, B. Zeng, and S. Liu, "Supervised homography learning with realistic dataset generation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 9806–9815.
- [116] D. Cattaneo, M. Vaghi, A. L. Ballardini, S. Fontana, D. G. Sorrenti, and W. Burgard, "Cmrnet: Camera to lidar-map registration," in *2019 IEEE intelligent transportation systems conference (ITSC)*. IEEE, 2019, pp. 1283–1289.
- [117] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 3354–3361.
- [118] X. Huang, P. Wang, X. Cheng, D. Zhou, Q. Geng, and R. Yang, "The apolloscape open dataset for autonomous driving and its application," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 10, pp. 2702–2719, 2019.
- [119] W. Lu, Y. Zhou, G. Wan, S. Hou, and S. Song, "L3-net: Towards learning based lidar localization for autonomous driving," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6389–6398.
- [120] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Lioung, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11621–11631.
- [121] J. Mao, M. Niu, C. Jiang, X. Liang, Y. Li, C. Ye, W. Zhang, Z. Li, J. Yu, C. Xu *et al.*, "One million scenes for autonomous driving: Once dataset," 2021.
- [122] H. Yu, Y. Luo, M. Shu, Y. Huo, Z. Yang, Y. Shi, Z. Guo, H. Li, X. Hu, J. Yuan *et al.*, "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*, 2022, pp. 21361–21370.
- [123] H. Li, K. Luo, B. Zeng, and S. Liu, "Gyroflow+: Gyroscope-guided unsupervised deep homography and optical flow learning," *arXiv preprint arXiv:2301.10018*, 2023.