Deep Learning-Based Calibration for a Two-Stage AVM Framework

1st Shin Jae Kang
School of Electronic and Electrical Engineering
Graduate School, Kyungpook National University
Daegu, Republic of Korea
kangsj129@knu.ac.kr

2nd Dong Seog Han
School of Electronic and Electrical Engineering
Graduate School, Kyungpook National University
Daegu, Republic of Korea
dshan@knu.ac.kr

3rd Han Gu Kim
School of Electronic Engineering
Undergraduate School, Kyungpook National University
Daegu, Republic of Korea
khangu0729@knu.ac.kr

***Abstract*—Surround-view monitoring systems for heavy-duty vehicles must maintain reliable BEV (bird's-eye-view) imagery over long-term operation in harsh environments. Classical geometric calibration with checkerboards or ChArUco boards provides accurate intrinsics and extrinsics at installation time, but cannot easily cope with parameter drift caused by vibration, thermal cycles, and minor impacts. This paper proposes a two-stage AVM (around-view monitoring) framework that combines pattern-based commissioning with learning-based maintenance of camera intrinsics. A learning-based calibration network predicts a ray-direction field from natural driving scenes and fits a parametric fisheye model, enabling both user-initiated recalibration and automatic monitoring of intrinsics drift during runtime. Experiments on a four-camera surround-view platform show that the proposed method reduces reprojection error from 102.44/107.52 pixels (GeoCalib$_{\text{radial/gen}}$) to 47.304 pixels (Ours$_{\text{dist}}$), lowers angular error from 4.28/7.24° to 0.77°, and decreases vertical/horizontal field-of-view errors from 7.45/14.88° to 0.49/1.26°, while matching or improving the accuracy of recent learning-based calibration. These gains translate into visibly more stable surround-view images compared to purely geometric baselines.**

***Index Terms*—Surround-view monitoring, fisheye camera, learning-based calibration, bird's-eye view, autonomous driving.**

I. Introduction

Surround-view monitoring is a key safety function for heavy-duty and special-purpose vehicles operating in narrow depots, construction sites, and logistics hubs. Drivers rely on stitched BEV images to avoid collisions with nearby obstacles and workers, but long-term field operation shows that AVM image quality gradually deteriorates: seams open up, objects become misaligned, and operators lose trust in the system. This degradation is caused by accumulated physical changes—vibration, minor impacts, temperature cycles, humidity, and cable strain—that perturb camera intrinsics and extrinsics, so even pixel-level parameter drift produces visible artifacts and safety-critical perception errors.

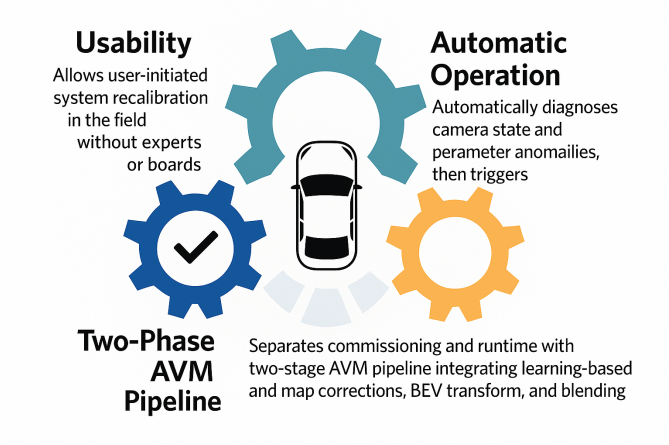

Usability
Allows user-initiated system recalibration in the field without experts or boards
Automatic Operation
Automatically diagnoses camera state and perameter anomailies, then triggers
Two-Phase AVM Pipeline
Separates commissioning and runtime with two-stage AVM pipeline integrating learning-based and map corrections, BEV transform, and blending

Fig. 1. High-level design goals of the proposed AVM system: usability via field recalibration, automatic operation via self-diagnosis and correction, and a two-stage AVM pipeline separating commissioning and runtime.

Classical geometric calibration with checkerboards or ChArUco boards remains the de-facto standard for setting up AVM systems and can achieve sub-pixel reprojection error. However, it requires vehicle downtime, controlled space, calibration patterns with sufficient coverage, and expert operation, so in practice it is performed only at installation or major maintenance while the system is expected to operate for years in harsher conditions [1], [2]. This gap motivates learning-based calibration that works on natural scenes, monitors the consistency between live images and the stored camera model, and enables in-field adjustment [3], [4]. In this context, calibration should be not only accurate but also *practically usable and capable of automatic operation*. Fig. 1 summarizes these three design goals in terms of usability, automatic operation, and a two-phase AVM pipeline.

To realize this goal, we formulate AVM operation as a two-stage pipeline from initial commissioning to long-term runtime. During commissioning, geometric tools estimate baseline

Identify applicable funding agency here. If none, delete this.

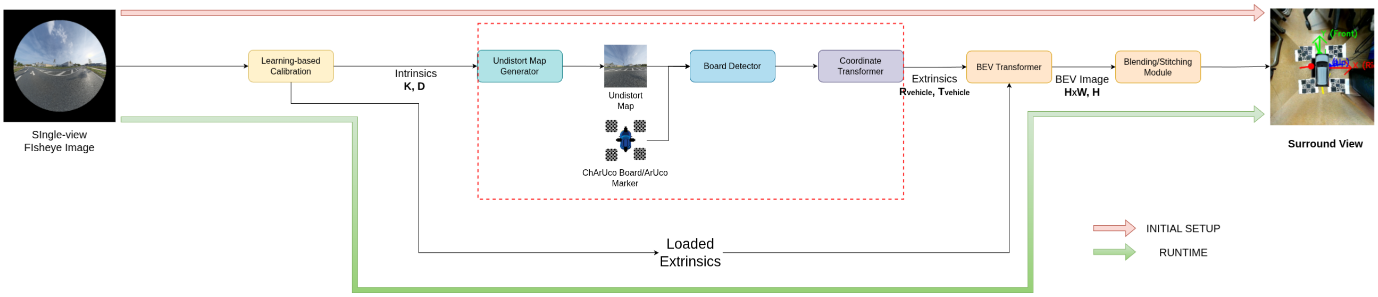


Fig. 2. Two-stage AVM pipeline. During commissioning (red path), learning-based calibration estimates intrinsics, pattern-based tools estimate extrinsics, and a BEV transformer plus blending module generate a reference surround-view. During runtime (green path), stored extrinsics are reused while the learning-based module updates effective intrinsics and maintains BEV quality.

intrinsics and extrinsics and generate a reference surround-view image [5], [6]. During runtime, the stored extrinsics are reused while a learning-based calibration module analyzes live fisheye images, detects intrinsics drift, and maintains BEV image quality under everyday disturbances, combining high-precision pattern-based calibration with continuous learning-based maintenance.

Based on this perspective, this paper makes the following contributions:

- **Learning-based calibration for usability.** We introduce a learning-based calibration model that operates on natural driving scenes so that users can re-check and refine camera parameters without dedicated patterns or expert tools, improving the usability and maintainability of AVM in the field.
- **Learning-based calibration for automatic operation.** We extend the same model into a runtime diagnostic module that monitors the consistency between live fisheye images and the stored camera model, detects intrinsics drift, and triggers automatic correction, enabling system-level "automatic operation" in which the AVM can self-diagnose and self-calibrate.
- **Two-stage AVM framework.** We integrate the proposed module into a two-stage AVM framework that separates commissioning and runtime, and we experimentally show that this design maintains surround-view image quality more robustly than conventional pattern-only calibration.

II. Proposed AVM Framework

A. Two-Stage AVM Pipeline

Fig. 2 gives an overview of the proposed AVM pipeline, which explicitly separates an *initial commissioning* path from a *runtime* path. During commissioning, a single-view fisheye image from each camera is first fed to a learning-based calibration module that estimates effective intrinsics $(\mathbf{K}, \mathbf{D})$ for the given hardware and mounting configuration. These intrinsics are passed to an undistort-map generator, which produces a dense rectification map and a rectified image for each camera. On the rectified views, a board detector locates checkerboard or ChArUco patterns [2], [7], and a coordinate transformer then estimates extrinsics $(\mathbf{R}_{\text{vehicle}}, \mathbf{t}_{\text{vehicle}})$ that anchor all cameras to the vehicle frame.

Given the estimated intrinsics and extrinsics, the system computes a BEV transform and feeds the resulting BEV images into a deterministic blending/stitching module [8], [9]. This module is responsible for composing the multi-camera BEV into a single surround-view image while suppressing seams, ghosting, and geometric inconsistency. The resulting parameters and maps—intrinsics, extrinsics, and BEV warps—are stored as the baseline calibration and are reused during runtime.

At runtime, the same extrinsics are loaded directly without requiring boards or vehicle downtime. Live fisheye images are continuously processed by the learning-based calibration module, which tracks potential intrinsics drift caused by vibration, temperature cycles, or minor impacts. When drift is detected, the module updates the effective intrinsics while keeping the extrinsics fixed, and the downstream undistort, BEV transform, and blending steps are updated accordingly. In this way, the proposed framework combines the high accuracy of pattern-based commissioning with continuous, learning-based maintenance during normal operation.

B. Learning-Based Calibration Architecture

The internal structure of the learning-based calibration module is illustrated in Fig. 3. A single fisheye input image is first processed by a convolutional feature extractor based on a lightweight ResNet backbone [10]. This stage converts the raw image into a multi-scale feature tensor that captures local texture and edge information while preserving the global layout of the scene.

The extracted feature maps are then tokenized and passed to a Vision Transformer (ViT) encoder [11]. The encoder models long-range dependencies across the full field of view, which is critical for fisheye images where rays near the periphery correspond to large angular changes. Through self-attention, the ViT aggregates evidence across the entire image and produces a compact latent representation that encodes both appearance and underlying ray geometry.

A shallow CNN decoder maps this latent representation back to a dense latent field defined over the image plane. From this latent field, a small prediction head regresses per-pixel or low-dimensional ray-vector descriptors, which are subsequently interpreted by a closed-form camera model to

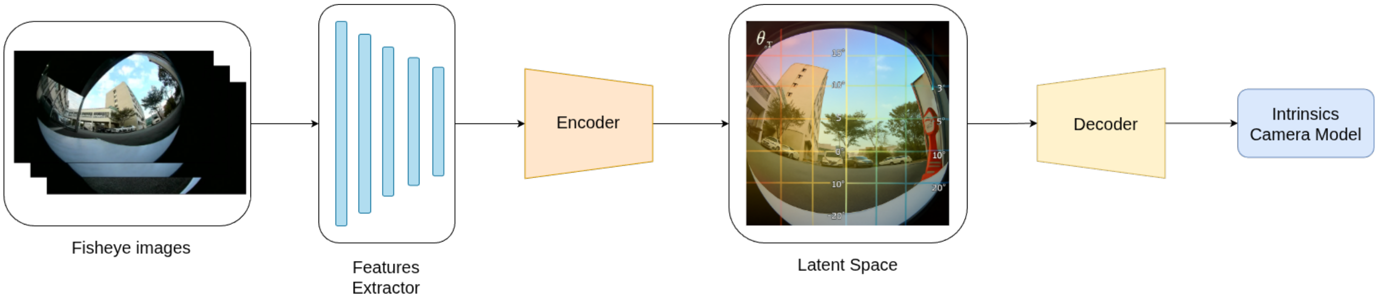


Fig. 3. Learning-based calibration architecture. A ResNet feature extractor produces image features, which are encoded by a ViT and decoded by a lightweight CNN into a latent field. From this field, ray-vector descriptors are inferred and fitted with a parametric fisheye camera model to obtain intrinsics $(\mathbf{K}, \mathbf{D})$.

obtain intrinsics parameters such as focal length, principal point, and distortion coefficients. In other words, the network learns to predict a ray-direction field in the latent space, and the camera model fits a parametric fisheye projection that best explains this field, similar in spirit to AnyCalib [12].

Because the model operates on ordinary driving scenes, it can be applied repeatedly during runtime without any calibration pattern [13], [14]. When the predicted intrinsics deviate from the stored commissioning values beyond a threshold, the system treats this as evidence of parameter drift and updates the effective intrinsics used by the undistort map and BEV transformer. This design allows the same architecture to support both user-initiated recalibration in the field and fully automatic monitoring and correction of camera intrinsics.

III. Experiments

A. Experimental Setup

We evaluate the proposed learning-based calibration on a surround-view system installed on a vehicle platform with four fisheye cameras mounted at the front, rear, left, and right. Each camera captures 1920×1080 images at 30 FPS. From continuous video, we extract a dataset of 2,000 single frames containing a mixture of outdoor and indoor scenes, building facades, vehicles, hand-drawn lines, and geometric patterns.

For surround-view generation, all methods use the same BEV configuration: a 1024×1024 canvas corresponding to roughly 50 pixels/m, and a fixed set of extrinsics that map each camera to the vehicle coordinate frame. The AVM pipeline runs on an Ubuntu environment.

To provide a reliable reference, we first perform a conventional board-based geometric calibration with checkerboard/ChArUco targets. The resulting intrinsics and extrinsics are treated as a pseudo ground truth (GT) for the given hardware configuration. Our learning-based method is trained using only images, but all quantitative comparisons are reported with respect to the board-based calibration and the surround-view images generated from its parameters.

B. Optimization of Camera Model

Before training the final network, we study which projection model is most suitable for our fisheye cameras. Fig. **??** plots, for each candidate model, the mapping from incident angle θ to image radius $r(\theta)$ for the ground-truth parameters and the parameters predicted by learning-based optimization. The five subplots correspond to UCM, EUCM, pinhole, edit, and dist models, arranged from the top left to the bottom right.

TABLE I
Calibration accuracy for different methods. Lower is better for all metrics.

Method	RE [pix]↓	AE [°]↓	vFoV err [°]↓	hFoV err [°]↓
GeoCalib$_{\text{radial}}$ [17]	102.44	4.28	7.45	14.88
GeoCalib$_{\text{gen}}$ [17]	107.52	7.24	9.59	11.01
AnyCalib$_{\text{dist}}$ [12]	48.764	**0.77**	0.51	**1.26**
AnyCalib$_{\text{gen}}$ [12]	51.498	1.04	0.52	**1.26**
Ours$_{\text{dist}}$	**47.304**	**0.77**	**0.49**	**1.26**
Ours$_{\text{gen}}$	51.521	1.04	**0.49**	**1.26**

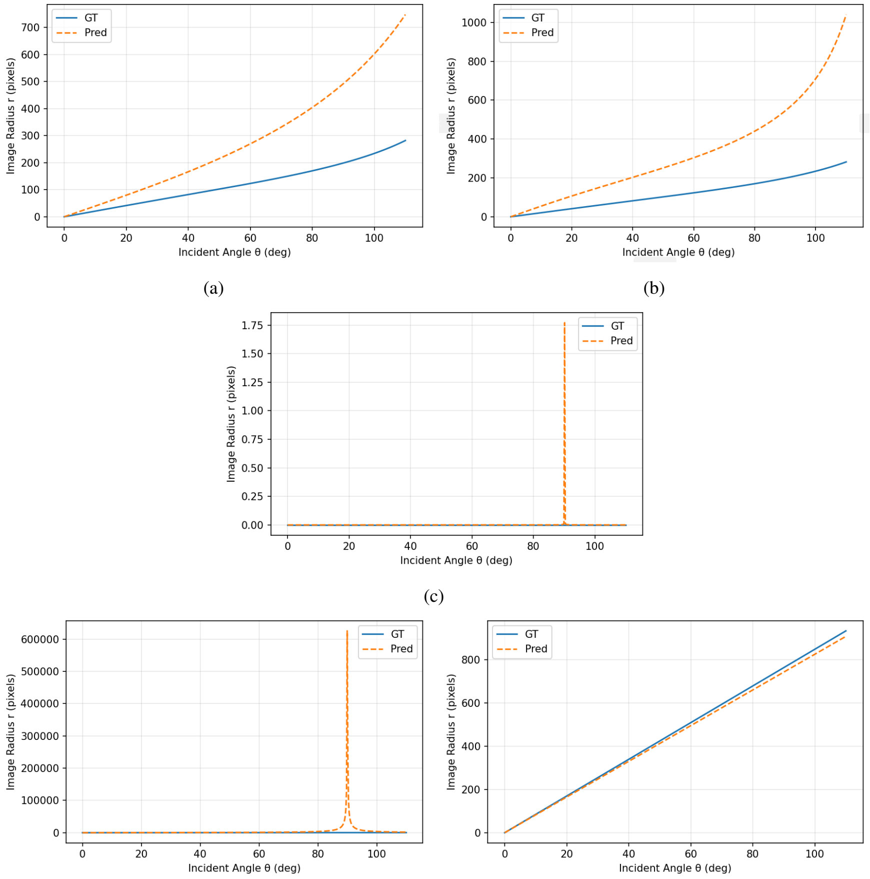
For UCM and EUCM, the predicted curves deviate significantly from the ground truth, especially at large incident angles. The radius grows too quickly near the periphery, producing strong over-stretching and poor extrapolation beyond the range covered by calibration data. The pinhole model behaves similarly and cannot reproduce the heavy fisheye distortion. The edited model gains flexibility but exhibits unstable behavior: the radius curve shows a sharp spike or collapse around high angles, which leads to severe artifacts in undistorted images.

In contrast, the dist model closely matches the ground-truth curve across the entire field of view. The GT and predicted curves almost overlap, with no sudden divergence or singular behavior. This indicates that the dist parameterization provides enough degrees of freedom to fit real fisheye lenses while remaining numerically stable for learning. Based on this observation, we adopt the dist camera model as the default projection model in all subsequent experiments.

C. Quantitative Results and Analysis

Table I reports numerical errors for different calibration pipelines in terms of pixel reprojection error (RE), angular error (AE), and vertical/horizontal field-of-view (vFoV/hFoV) deviations.

GeoCalib$_{\text{radial}}$ and GeoCalib$_{\text{gen}}$ denote our baseline learning-based calibration variants configured with a radial distortion model and a more generic wide-angle model, respectively.



(a)

(b)

(c)

(d)

(e)

Fig. 4. Comparison of radius–angle mappings between the board-calibrated ground truth and the learned camera model. (a) UCM. (b) EUCM. (c) Pinhole. (d) EDiT. (e) Dist. Blue curves represent the board-based ground truth, while orange curves indicate the learned model. Among the tested parameterizations, the Dist model exhibits the closest agreement with the ground truth across the full field of view.

They show the largest errors in this benchmark, with RE above 100 pixels and FoV deviations exceeding $7°$–$14°$, which indicates that these parameterizations are not well suited to our fisheye surround-view setup. AnyCalib$_{\text{dist}}$ and AnyCalib$_{\text{gen}}$ already reduce RE by more than half (down to 48.764 and 51.498 pixels, respectively) and significantly improve AE as well as vFoV/hFoV accuracy. Our method further refines the camera model on top of these learning-based baselines: Ours$_{\text{dist}}$ achieves the lowest RE of 47.304 pixels while keeping AE at $0.77°$ and reducing vFoV error to $0.49°$ with an hFoV error of only $1.26°$. The generic variant Ours$_{\text{gen}}$ matches the FoV accuracy of Ours$_{\text{dist}}$ (vFoV $0.49°$, hFoV $1.26°$) and closely tracks its RE and AE. Overall, the proposed calibration scheme improves upon the baseline GeoCalib variants and reaches or surpasses AnyCalib across all metrics.

D. *Qualitative Results and Analysis*

Fig. 5 shows qualitative surround-view results for the geometric baseline and the proposed method. With purely geometric calibration, the stitched BEV image exhibits visible seams and distortions around the vehicle. Objects near the overlapping regions appear slightly stretched or duplicated, and the rectangular platform in the center of the scene is warped, with its edges misaligned across camera boundaries.

Using the proposed learning-based calibration, the surround-view image becomes noticeably more coherent. The rectangular platform and nearby shelves appear more regular and symmetric; seams are less noticeable, and high-contrast structures such as lines on the floor are better aligned across views. However, minor residual distortions remain in some far-range areas. These artifacts mainly arise from accumulated errors in converting each camera's extrinsics to the unified vehicle coordinate frame via the multi-step bridge transformation [15], [16]. Even small pose errors in that chain can be amplified in the BEV domain. Despite this limitation, the qualitative comparison clearly demonstrates that the proposed calibration improves the perceptual stability and usability of the AVM image over the geometric baseline.

IV. Conclusion

This paper presented a two-stage surround-view framework that augments classical board-based calibration with learning-based maintenance of fisheye intrinsics. By combining a learning-based calibration network with a stable dist camera model, the proposed method can operate on natural scenes,

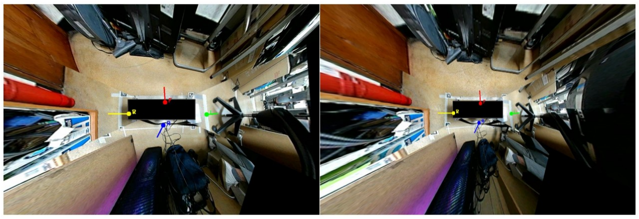


Fig. 5. Qualitative surround-view results. Left: geometric calibration with board-based intrinsics only. Right: proposed learning-based calibration (dist model). The proposed method reduces seams and distortion around the vehicle, while small residual artifacts remain due to accumulated errors in extrinsics to vehicle-frame conversion.

enabling user-initiated recalibration and automatic monitoring of intrinsics drift during runtime. Experiments on a four-camera AVM platform showed that the method matches or surpasses recent learning-based calibration in reprojection and field-of-view errors, while also delivering qualitatively more coherent surround-view images. In future work, we plan to extend the approach to joint refinement of intrinsics and extrinsics, and to integrate uncertainty estimates so that the AVM system can reason about calibration confidence in real time.

Acknowledgment

This study was supported by the BK21 Four project funded by the Ministry of Education, Korea (4199990113966)

References

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

[2] S. Garrido-Jurado, R. Muñoz-Salinas, F. J. Madrid-Cuevas, and M. J. Marín-Jiménez, "Automatic generation and detection of highly reliable fiducial markers under occlusion," *Pattern Recognition*, vol. 47, no. 6, pp. 2280–2292, 2014.

[3] O. Bogdan, V. Badrinarayanan, and R. Cipolla, "DeepCalib: a deep learning approach for automatic intrinsic calibration of wide field-of-view cameras," in *Proc. British Mach. Vis. Conf. (BMVC)*, 2018.

[4] A. Kendall, M. Grimes, and R. Cipolla, "PoseNet: A convolutional network for real-time 6-DOF camera relocalization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2015, pp. 2938–2946.

[5] J. Lee, J. Lim, D. Kim, S. Oh, and I.-S. Kweon, "Surround view camera calibration for ADAS systems," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, 2012, pp. 360–365.

[6] J. Choi, S. Kim, and K. Yun, "Automatic calibration of multi-camera surround-view systems," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 10, pp. 3820–3832, 2019.

[7] A. Romero-Ramirez, R. Muñoz-Salinas, and R. Medina-Carnicer, "Speeded up detection of squared fiducial markers," *Image and Vision Computing*, vol. 76, pp. 38–47, 2018.

[8] A. Broggi, P. Cerri, S. Debattisti, and M. C. Laghi, "A real-time multi-resolution approach to vision-based pedestrian detection," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, 2010, pp. 194–199.

[9] L. Güvenç, O. Çelik, and M. Akar, "Stitching and blending of bird's eye view images for driver assistance systems," in *Proc. IEEE Int. Conf. Mechatronics*, 2013, pp. 402–407.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770–778.

[11] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021.

[12] J. Valiente, P. Miraldo, and T.-J. Chin, "AnyCalib: Arbitrary camera model calibration from planar patterns," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 7190–7199.

[13] Y. Li, H. Huang, and J. Ren, "DeepCalib: A deep learning framework for automatic camera calibration," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, 2019, pp. 1905–1909.

[14] H. Huang, Y. Li, and S. Shen, "MetaCalib: Few-shot camera calibration using meta-learning," *IEEE Robot. Autom. Lett.*, vol. 6, no. 3, pp. 5163–5170, 2021.

[15] L. Chen, Z. Yang, Y. Chen, and K. Wang, "Multi-view fusion for bird's-eye-view reconstruction in autonomous driving," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA) Workshops*, 2017.

[16] M. Liang, B. Yang, S. Wang, and R. Urtasun, "Multi-task multi-sensor fusion for 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 7345–7353.

[17] A. Veicht, P.-E. Sarlin, P. Lindenberger, and M. Pollefeys, "GeoCalib: Learning Single-image Calibration with Geometric Optimization," arXiv preprint arXiv:2409.06704, 2024.