Table 1. Comparison of the features of conventional methods and our proposed method

Method		DL[1]	Heatmap[1]	Manhattan[1]	Pan	Tilt & Roll	Distortion	Projection
Non-Manhattan world								
López-Antequera *et al.* [33]	CVPR'19	✓				✓	✓	Perspective
Wakai and Yamashita [52]	ICCVW'21	✓				✓	✓	Equisolid angle
Wakai *et al.* [53]	ECCV'22	✓				✓	✓	Generic camera [53]
Manhattan world								
Wildenauer *et al.* [57]	BMVC'13			✓	✓	✓	✓	Division model [14]
Antunes *et al.* [3]	CVPR'17			✓	✓	✓	✓	Division model [14]
Pritts *et al.* [41]	CVPR'18			✓	✓	✓	✓	Division model [14]
Lochman *et al.* [32]	WACV'21			✓	✓	✓	✓	Division model [14]
Ours		✓	✓	✓	✓	✓	✓ [53]	Generic camera [53]

[1] DL is learning-based methods; Heatmap is heatmap regression; Manhattan is based on the Manhattan world for world coordinates

objects because these methods need to detect many arcs to estimate the VPs. Therefore, city scenes in which sky or street trees dominate the images degrade the performance of geometry-based methods.

On the basis of the observations above, to achieve accurate and robust estimation, we propose a learning-based calibration method that estimates extrinsics (pan, tilt, and roll angles), focal length, and a distortion coefficient simultaneously from a single image in Figure 1. Our heatmap regression estimates each direction using labeled image coordinates to distinguish the four directions of a road intersection in a Manhattan world. Furthermore, we introduce additional geometric keypoints, called auxiliary diagonal points (ADPs), to compensate for the lack of VPs in each image.

To investigate the effectiveness of the proposed methods, we conducted extensive experiments on three large-scale datasets [9, 38, 64] as well as off-the-shelf cameras. This evaluation demonstrated that our method notably outperforms conventional geometry-based [32, 41] and learning-based [33, 52, 53] methods. The major contributions of our study are summarized as follows:

- We propose a heatmap-based VP estimator for recovering the rotation from a single image to achieve higher accuracy and robustness than geometry-based methods using arc detectors.
- We introduce auxiliary diagonal points with an optimal 3D arrangement based on the spatial uniformity of regular octahedron groups to address the lack of VPs in an image.

2. Related work

Camera model. For geometric tasks, camera calibration estimates the parameters in a camera model. This model expresses a mapping from world coordinates $\tilde{\mathbf{p}}$ to image coordinates $\tilde{\mathbf{u}}$ in homogeneous coordinates. This mapping is conducted using extrinsic and intrinsic parameters. Extrinsic parameters $[\,\mathbf{R} \mid \mathbf{t}\,]$ consist of a rotation matrix $\mathbf{R}$ and a translation vector $\mathbf{t}$ to represent the relation between the origins of the camera coordinates and Manhattan world coordinates (or other world coordinates). The intrinsic parameters are distortion γ, image sensor pitch (d_u, d_v), and a principal point (c_u, c_v). The subscripts u and v indicate the horizontal and vertical directions, respectively. The mapping is formulated as

$$\tilde{\mathbf{u}} = \begin{bmatrix} \gamma/d_u & 0 & c_u \\ 0 & \gamma/d_v & c_v \\ 0 & 0 & 1 \end{bmatrix} [\,\mathbf{R} \mid \mathbf{t}\,]\, \tilde{\mathbf{p}}. \qquad (1)$$

Kannala and Brandt [16] proposed the generic camera model, which includes fisheye lens cameras and is given by

$$\gamma = \tilde{k}_1 \eta + \tilde{k}_2 \eta^3 + \cdots, \qquad (2)$$

where $\tilde{k}_1, \tilde{k}_2, \ldots$ are distortion coefficients and η is an incident angle. Wakai *et al.* [53] proposed an alternative generic camera model for learning-based methods, expressed as

$$\gamma = f \cdot (\eta + k_1 \eta^3), \qquad (3)$$

where f is focal length and k_1 is a distortion coefficient. Although Equation (3) is only a third-order polynomial with respect to η, the model can practically express fisheye projection with sub-pixel error [53].

Manhattan world. Coughlan *et al.* [12] proposed the Manhattan world for human navigation on the basis of the prior over edge models. The Manhattan world assumption regards the world as consisting of grid-shaped roads; that is, two of the three orthogonal coordinate axes lie along a crossroads, and the remaining axis is vertical. Given the Manhattan world O_M-$X_M Y_M Z_M$ in Figure 2, camera angles are defined as a rotation matrix $\mathbf{R}$ that is compatible with pan, tilt, and roll angles. In this paper, we ignore the relations between the extrinsics of the camera and the body of cars, drones, or robots because these relations can be determined using designed values or calibration. Therefore, the task of camera calibration is to determine camera angles of a 3D-rotated camera in a Manhattan world.

Camera calibration. Perspective camera calibration methods in the Manhattan world have been proposed for Hough-transform-based methods [43, 44] and VP-based methods [8, 11, 18, 19, 22–24, 30, 35, 45, 46, 48, 49, 58, 63]. However, these methods address only narrow FOV cameras without distortion.