of a polyhedron. We cannot escape the trade-off between the strength of constraints and the ease of training. This trade-off depends on the arrangement of the directions of points and the number of directions.

To solve this problem concerning the arrangement and number of points, we define additional VP-related points called ADPs based on the spatial symmetry as follows. We found that six 3D VPs form a regular octahedron that has the symmetry of regular octahedron groups (octahedral symmetry), see Figure 3. This symmetry means that a regular octahedron has three types of rotational symmetric axes. To estimate a unique camera rotation requiring two or more directions, the point arrangement prefers 3D spatial uniformity, such as the vertexes of regular polygons. Considering the wide spatial uniformity and small number of points, we use the ADPs, which are the eight diagonal points that indicate the directions of the cubic corners in Table 2. This arrangement of VPs and ADPs (VP/ADPs) also has the symmetry of regular octahedron groups and the greatest spatial uniformity in the case of eight points because of the diagonal directions, as shown in Figure 3. Therefore, 3D VP/ADP coordinates are the optimal arrangement given a practical number of points. The supplementary materials describe details of the symmetry and optimal arrangement.

3.2. Network architecture

Vanishing-point estimator. We found that VP estimation in images corresponds to single human pose estimation [2] in terms of labeled image-coordinate detection. These two tasks, VP detection and pose estimation, are similar because of the geometric relations of VPs and pose keypoints; that is, 3D VPs form a regular octahedron, and pose keypoints are based on a human skeleton. This similarity suggests that heatmap regression can achieve accurate and robust VP estimation as it does for single human pose estimation. Furthermore, using ADPs to increase the number of points, we can overcome the problem of the heatmap regression for VPs; that is, a unique camera rotation cannot be determined for images with few VPs.

Additionally, Wakai *et al*. [53] reported that mismatches in dataset domains degrade calibration accuracy. This degradation suggests that regressors consisting of fully connected layers extract domain-specific features without geometric cues such as VPs. In contrast to conventional regressors, heatmap regressors using 2D Gaussian kernels on labeled points have the potential for pixel-wise accuracy in pose estimation [37]. Therefore, we propose a heatmap-regression network, called the "VP estimator," that detects image VP/ADPs and is likely to avoid such degradation.

Distortion estimator. To estimate camera rotation, we require intrinsic parameters that project image coordinates to 3D incident ray vectors because the rotation is calculated using these incident ray vectors in Manhattan world coordinates. For the intrinsics in Equation (3), we use Wakai *et al*.'s calibration network [53] without the tilt and roll angle regressors, which is called the "distortion estimator." Therefore, our network has two estimators in Figure 1.

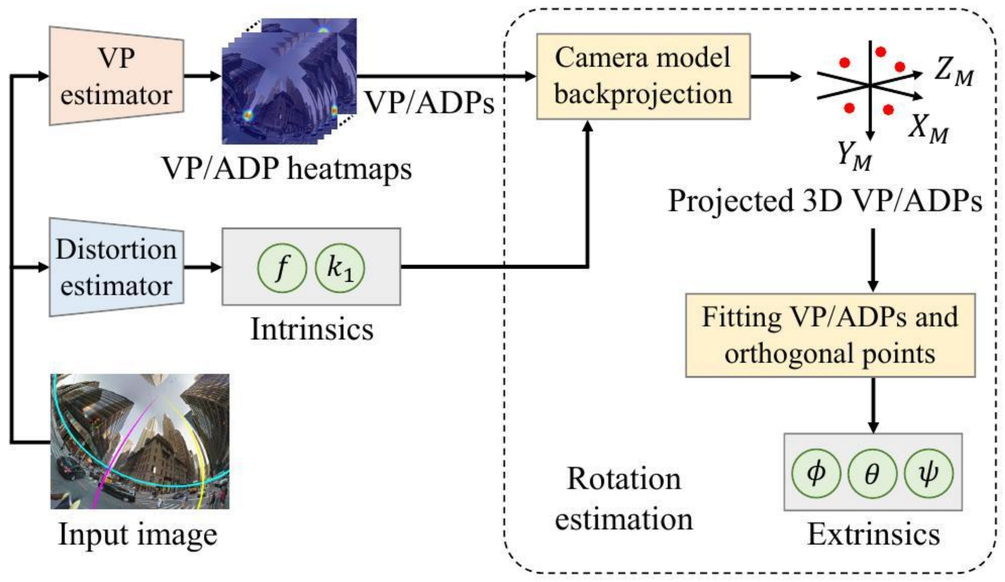

Figure 4. Calibration pipeline for inference. The intrinsics are estimated by the distortion estimator. Camera models project VP/ADPs onto the unit sphere using backprojection. The extrinsics are calculated from the fitting. The input fisheye image is generated from [38].

Implementation details. We use the HRNet [47] backbone, which has shown strong performance in various tasks, for our VP estimator. We found that the HRNet loss function evaluates only images that include detected keypoints; that is, detection failure does not affect the loss value. To tackle this problem, we modified this loss function to evaluate all images, including those with detection failure. This modification is suitable for deep single image camera calibration because, unlike human pose estimation, we always estimate camera rotation. To achieve sub-pixel precision, DARK [60] is applied to the heatmaps as postprocessing. Note that we do not employ DARK to generate heatmaps as a preprocessing step because such preprocessing leads to inconsistency between the heatmaps and camera parameters. Rectangular images are center cropped for the VP estimator.

3.3. Training and inference

Training. Using the generated fisheye images with ground-truth camera parameters in Section 4.1 and VP/ADP labels in Section 4.2, we train our two estimators independently. The VP estimator is trained using the modified HRNet loss function described above. In addition, the distortion estimator is trained using the harmonic non-grid bearing loss [53].

Inference. Figure 4 shows our calibration pipeline for the inference. First, we obtain the VP/ADPs from the VP estimator and intrinsics from the distortion estimator. Second, these 2D VP/ADPs are projected onto a unit sphere in world coordinates using backprojection. Finally, we convert the 3D VP/ADPs to the extrinsics, as described below.

Here, we describe the estimation of camera rotation from the VP/ADPs. Note that this estimation, which is based on