Tame a Wild Camera: In-the-Wild Monocular Camera Calibration

Shengjie Zhu, Abhinav Kumar, Masa Hu, and Xiaoming Liu
Department of Computer Science and Engineering, Michigan State University, East Lansing, MI, 48824
{zhusheng, kumarab6, huynshen}@msu.edu, liuxm@cse.msu.edu

Abstract

3D sensing for monocular in-the-wild images, e.g., depth estimation and 3D object detection, has become increasingly important. However, the unknown intrinsic parameter hinders their development and deployment. Previous methods for the monocular camera calibration rely on specific 3D objects or strong geometry prior, such as using a checkerboard or imposing a Manhattan World assumption. This work instead calibrates intrinsic via exploiting the monocular 3D prior. Given an undistorted image as input, our method calibrates the complete 4 Degree-of-Freedom (DoF) intrinsic parameters. First, we show intrinsic is determined by the two well-studied monocular priors: monocular depthmap and surface normal map. However, this solution necessitates a low-bias and low-variance depth estimation. Alternatively, we introduce the incidence field, defined as the incidence rays between points in 3D space and pixels in the 2D imaging plane. We show that: 1) The incidence field is a pixel-wise parametrization of the intrinsic invariant to image cropping and resizing. 2) The incidence field is a learnable monocular 3D prior, determined pixel-wisely by up-to-scale monocular depthmap and surface normal. With the estimated incidence field, a robust RANSAC algorithm recovers intrinsic. We show the effectiveness of our method through superior performance on synthetic and zero-shot testing datasets. Beyond calibration, we demonstrate downstream applications in image manipulation detection & restoration, uncalibrated two-view pose estimation, and 3D sensing. Codes/models are held here.

1 Introduction

Camera calibration is typically the first step in numerous vision and robotics applications [27, 46] that involve 3D sensing. Classic methods enable accurate calibration by imaging a specific 3D structure such as a checkerboard [49]. With the rapid growth of monocular 3D vision, there is an increasing focus on 3D sensing from in-the-wild images, such as monocular depth estimation, 3D object detection [35, 36], and 3D reconstruction [42, 43]. While 3D sensing techniques over in-the-wild images are developed, camera calibration for such in-the-wild images continues to pose significant challenges.

Classic methods for monocular calibration use strong geometry prior, such as using a checkerboard. However, such 3D structures are not always available in in-the-wild images. As a solution, alternative methods relax the assumptions. For example, [29] and [23] calibrate using common objects such as human faces and objects’ 3D bounding boxes. Another significant line of research [7, 19, 34, 37, 51, 62, 66] is based on the Manhattan World assumption [15], which posits that all planes within a scene are either parallel or perpendicular to each other. This assumption is further relaxed [28, 30, 64] to estimate the lines that are either parallel or perpendicular to the direction of gravity. The intrinsic parameters are recovered by determining the intersected vanishing points of detected lines, assuming a central focal point and an identical focal length.

Figure 1: In (a), our work focuses on monocular camera calibration for in-the-wild images. We recover the intrinsic from monocular 3D-prior. In (c) - (e), an estimated depthmap is converted to surface normal using a groundtruth and noisy intrinsic individually. Noisy intrinsic distorts the point cloud, consequently leading to inaccurate surface normal. In (e), the normal presents a different color to (c). Motivated by the observation, we develop a solver that utilizes the consistency between the two to recover the intrinsic. However, this solution exhibits numerical instability. We then propose to learn the incidence field as an alternative 3D monocular prior. The incidence field is the collection of the pixel-wise incidence ray, which originates from a 3D point, targets at a 2D pixel, and crosses the camera origin, as shown in (b). Similar to depthmap and normal, a noisy intrinsic leads to a noisy incidence field, as in (e). By same motivation, we develop neural network to learn in-the-wild incidence field and develop a RANSAC algorithm to recover intrinsic from the estimated incidence field.

While the assumptions are relaxed, they may still not hold true for in-the-wild images. This creates a contradiction: although we develop robust models to estimate in-the-wild monocular depthmap, generating its 3D point cloud remains infeasible due to the missing intrinsic. A similar challenge arises in monocular 3D object detection, as we face limitations in projecting the detected 3D bounding boxes onto the 2D image. In AR/VR applications, the absence of intrinsic precludes placing multiple reconstructed 3D objects within a canonical 3D space. The absence of a reliable, assumption-free monocular intrinsic calibrator has become a bottleneck in deploying these 3D sensing applications.

Our method is motivated by the consistency between the monocular depthmap and surface normal map. In Fig. 1 (c) - (e), an incorrect intrinsic distorts the back-projected 3D point cloud from the depthmap, resulting in distorted surface normals. Based on this, intrinsic is optimal when the estimated monocular depthmap aligns consistently with the surface normal. We present a solution to recover the complete 4 DoF intrinsic by leveraging the consistency between the surface normal and depthmap. However, the algorithm is numerically ill-conditioned as its computation depends on the accurate gradient of depthmap. This requires depthmap estimation with low bias and variance.

To resolve it, we propose an alternative approach by introducing a novel 3D monocular prior in complementation to the depthmap and surface normal map. We refer to this as the incidence field, which depicts the incidence ray between the observed 3D point and the projected 2D pixel on the imaging plane, as shown in Fig. 1 (b). The combination of the incidence field and the monocular depthmap describes a 3D point cloud. Compared to the original solution, the incidence field is a direct pixel-wise parameterization of the camera intrinsic. This implies that a minimal solver based on the incidence field only needs to have low bias. We then utilize a deep neural network to perform the incidence field estimation. A non-learning RANSAC algorithm is developed to recover the intrinsic parameters from the estimated incidence field.

We consider the incidence field is also a monocular 3D prior. Similar to depthmap and surface normal, the incidence field is invariant to the image cropping or resizing. This encourages its generalization over in-the-wild images. Theoretically, we show the incidence field is pixel-wisely determined by depthmap and normal map. Empirically, to support our argument, we combine multiple public datasets into a comprehensive dataset with diverse indoor, outdoor, and object-centric images captured by different imaging devices. We further boost the variety of intrinsic by resizing and cropping the images in a similar manner as [23]. Finally, we include zero-shot testing samples to benchmark real-world monocular camera calibration performance.

We showcase downstream applications that benefit from monocular camera calibration. In addition to the aforementioned 3D sensing tasks, we present two intriguing additional applications. One is detecting and restoring image resizing and cropping. When an image is cropped or resized, it disrupts the assumption of a central focal point and identical focal length. Using the estimated
Table 1: Camera Calibration Methods from Strong to Relaxed Assumptions. Non-learning methods [23, 29, 34, 37, 38, 51, 62, 71–74] rely on strong assumptions. Learning-based methods [28, 30, 38] relax the assumptions to using gravity-aligned panorama images during training. Our method makes no assumptions except for undistorted images. This enables training with any calibrated images and calibrates 4 DoF intrinsic.

<table>
<thead>
<tr>
<th>Assumption</th>
<th>DoF</th>
<th>Specific-Objects</th>
<th>Manhattan</th>
<th>Train with Panorama Images</th>
<th>Undistorted Images</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-learning</td>
<td>4</td>
<td>[23, 29, 34, 37, 51, 62, 71–74]</td>
<td></td>
<td>[28, 30, 38]</td>
<td></td>
</tr>
<tr>
<td>Learning-based</td>
<td>1</td>
<td></td>
<td>[28, 30, 38]</td>
<td>[23, 28, 30, 38]</td>
<td></td>
</tr>
<tr>
<td>Ours</td>
<td>4</td>
<td></td>
<td></td>
<td>[23, 29, 38, 71–74]</td>
<td></td>
</tr>
</tbody>
</table>

Intrinsic parameters, we restore the edited image by adjusting its intrinsic to a regularized form. The other application involves two-view uncalibrated camera pose estimation. With established image correspondence, a fundamental matrix [25] is determined. However, there does not exist an injective mapping between the fundamental matrix and camera pose [27]. This raises a counter-intuitive fact: inferring the pose from two uncalibrated images is infeasible. But our method enables uncalibrated two-view pose estimation by applying monocular camera calibration.

We summarize our contributions as follows:
- Our approach tackles monocular camera calibration from a novel perspective by relying on monocular 3D priors. Our method makes no assumption for the to-be-calibrated undistorted image.
- Our algorithm provides robust monocular intrinsic estimation for in-the-wild images, accompanied by extensive benchmarking and comparisons against other baselines.
- We demonstrate its benefits on diverse and novel downstream applications.

2 Related Works

Monocular Camera Calibration with Geometry. One line of work [7, 19, 34, 37, 51, 62, 66] assumes the Manhattan World assumption [15], where all planes in 3D space are either parallel or perpendicular. Under the assumption, line segments in the image converge at the vanishing points, from which the intrinsic is recovered. LSD [61] and EDLine [2] develop robust line estimators. Others jointly estimate the horizon line and the vanishing points [40, 54, 70]. In Tab. 1, recent learning-based methods [28, 30, 38, 63, 64] relax the assumption by using the gravity-aligned panorama images during training. Such data provides the groundtruth vanishing point and horizon lines which are originally computed with the Manhattan World assumption. Still, the assumption constrains [28, 38, 63, 64] in modeling intrinsic as 1 DoF camera. Recently, [30] relaxes the assumption to 3 DoF via regressing the focal point. In comparison, our method makes no assumption except for undistorted images. This enables us to calibrate 4 DoF intrinsic and train with any calibrated images.

Monocular Camera Calibration with Object. Zhang’s method [73] based on a checkerboard pattern is widely regarded as the standard for camera calibration. Several works generalize this method to other geometric patterns such as 1D objects [74], line segments [72], and spheres [71]. Recent works [29] and [23] extend camera calibration to real-world objects such as human faces. Optimizers, including BPnP [12] and PnP [56] are developed. However, the usage of specific objects restricts their applications. In contrast, our approach is applicable to any image.

Image Cropping and Resizing. Detecting photometric image manipulation [5, 17, 24] is extensively researched. But few studies detecting geometric manipulation, e.g., resizing and cropping. On resizing, [18] regresses the image aspect ratio with a deep model. On cropping, a proactive method [67] is developed. We demonstrate that calibration also detects geometric manipulation. Our method does not need to encrypt images, complementing photometric manipulation detection.

Uncalibrated Two-View Pose Estimation. With the fundamental matrix estimated, the two-view camera pose is determined up to a projective ambiguity if images are uncalibrated. Alternative solutions [31, 47, 58] exist by employing deep networks to regress the pose. However, regression hinders the usage of geometric constraints, which proves crucial in calibrated two-view pose estimation [60, 75, 77]. Other works [20, 26] use more than two uncalibrated images for pose estimation. Our work complements prior studies by enabling an uncalibrated two-view solution.

Learnable Monocular 3D Priors. Monocular depth [39, 76] and surface normals [6] are two established 3D priors. Numerous studies have shown a learnable mapping between monocular images and corresponding 3D priors. Recently, [30] introduces the perspective field as a monocular 3D
In this section, we explain how to determine the intrinsic matrix $K$ using monocular 3D priors, such as the surface normal map and monocular depthmap. We then introduce the incidence field and show it a learnable monocular 3D prior for inferring gravity direction. In this work, we suggest the incidence field is also a learnable monocular 3D prior. A common characteristic of monocular 3D priors is their exclusive relationship with 3D structures, making them invariant to 2D manipulations, e.g., cropping, and resizing. In contrast, camera intrinsic changes on spacial manipulation. Unlike other monocular priors, computing the groundtruth incidence field is straightforward, relying solely on intrinsic and image coordinates.

### 3 Method

Our work focuses on intrinsic estimation from in-the-wild monocular images captured by modern imaging devices. Hence we assume an undistorted image as input. In this section, we first show how to estimate intrinsic parameters by using monocular 3D priors, such as the surface normal map and monocular depthmap. We then introduce the incidence field and show it a learnable monocular 3D prior invariant to 2D spacial manipulation. Next, we describe the training strategy and the network used to learn the incidence field. After estimating the incidence field, we present a RANSAC algorithm to recover the 4 DoF intrinsic parameters. Lastly, we explore various downstream applications of the proposed algorithm. Fig. 2 shows the framework of the proposed algorithm.

#### 3.1 Monocular Intrinsic Calibration

Our method aims to use generalizable monocular 3D priors without assuming the 3D scene geometry. Hence, we start with monocular depthmap $D$ and surface normal map $N$. Assume there exists a learnable mapping from the input image $I$ to depthmap $D$ and normal map $N$: $D, N = D_0(I)$, where $D_0$ is a learned network. We denote the intrinsic $K_{\text{simple}}$, $K$, and its inverse $K^{-1}$ as:

$$
K_{\text{simple}} = \begin{bmatrix} f & 0 & w/2 \\ 0 & f & h/2 \\ 0 & 0 & 1 \end{bmatrix}, \quad K = \begin{bmatrix} f_x & 0 & b_x \\ 0 & f_y & b_y \\ 0 & 0 & 1 \end{bmatrix}, \quad K^{-1} = \begin{bmatrix} 1/f_x & 0 & -b_x/f_x \\ 0 & 1/f_y & -b_y/f_y \\ 0 & 0 & 1 \end{bmatrix}. \quad (1)
$$

The notation $K_{\text{simple}}$ suggests a simple camera model with the identical focal length and central focal point assumption. Given a 2D homogeneous pixel location $p = [x \ y \ 1]^T$ and its depth value $d = D(p)$, the corresponding 3D point $P = [X \ Y \ Z]^T$ is defined as:

$$
P = \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = d \cdot K^{-1} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = d \cdot \begin{bmatrix} x - b_x/y \\ y - b_y/x \\ 1 \end{bmatrix} = d \cdot \begin{bmatrix} u_x \\ u_y \\ 1 \end{bmatrix} = d \cdot v,
$$

where the vector $v$ is an incidence ray, originating from the 3D point $P$, directed towards the 2D pixel $p$, and passing through the camera’s origin. The incidence field $V$ is the collection of incidence rays $v$ associated with all pixels at location $p$, where $v = V(p)$.

#### 3.2 Monocular Intrinsic Calibration with Monocular Depthmap and Surface Normal

In this section, we explain how to determine the intrinsic matrix $K$ using the estimated surface normal map $N$ and depthmap $D$. Given the estimated depth $d = D(p)$ and normal $n = N(p)$ at 2D pixel
location $p$, a local 3D plane is described as:

$$\mathbf{n}^\top \cdot d \cdot \mathbf{v} + c = 0.$$  \hfill (3)

By taking derivative in $x$-axis and $y$-axis directions, we have:

$$\mathbf{n}^\top \nabla_x (d \cdot \mathbf{v}) = 0, \quad \mathbf{n}^\top \nabla_y (d \cdot \mathbf{v}) = 0.$$  \hfill (4)

Note the bias $c$ of the 3D local plane is independent of the camera projection process. Without loss of generality, we show the case of our method for $x$-direction. Expanding Eq. (4), we obtain:

$$n_1 \nabla_x (d \cdot \frac{x - b_x}{f_x}) + n_2 \frac{y - b_y}{f_y} \nabla_x (d) + n_3 \nabla_x (d) = 0,$$  \hfill (5)

where $\nabla_x (d)$ represents the gradient of the depthmap $D$ in the $x$-axis and can be computed, for example, using a Sobel filter [33]. Next, re-parametrize the unknowns in Eq. (5) to get:

$$a_1 f_x f_y + a_2 f_x b_y + a_3 f_y b_x + a_4 f_y + a_5 f_x = 0.$$  \hfill (6)

Divide both sides of the equation by $f_x$ to get:

$$a_1 f_y + a_2 b_y + a_3 r b_x + a_4 r + a_5 = 0,$$  \hfill (7)

where $r = \frac{f_y}{f_x}$. By stacking Eq. (7) with $N \geq 4$ randomly sampled pixels, we acquire a linear system:

$$\mathbf{A}_{N \times 4} \mathbf{X}_{4 \times 1} = \mathbf{B}_{4 \times 1},$$  \hfill (8)

where the intrinsic parameter to be solved is stored in a vector $\mathbf{X}_{4 \times 1} = [f_y \ b_y \ \frac{r b_x}{r} \ r]^\top$. This solves the other intrinsic parameters as:

$$f_y = f_y, \quad b_y = b_y, \quad f_x = \frac{f_y}{r}, \quad b_x = \frac{r b_x}{r}.$$  \hfill (9)

The known constants are stored in matrix $\mathbf{A}_{N \times 4}$ and $\mathbf{B}_{4 \times 1}$. If we choose $N = 4$ in Eq. (8), we obtain a minimal solver where the solution $\mathbf{X}$ is computed by performing Gauss-Jordan Elimination. Conversely, when $N > 4$, the linear system is over-determined, and $\mathbf{X}$ is obtained using a least squares solver. The above suggests the intrinsic is recoverable from the monocular 3D prior.

### 3.3 Monocular Incidence Field as Monocular 3D Prior

Eq. (9) relies on the consistency between the surface normal and depthmap gradient, which may require a low-variance depthmap estimate. From Fig. 3, even groundtruth depthmap leads to spurious normal due to its inherent high variance. Thus, minimal solver in Eq. (9) can lead to a poor solution.
As a solution, we propose to directly learn the incidence field $V$ as a monocular 3D prior. In Eq. (2) and Fig. 1, the combination of the incidence field $V$ and the monocular depthmap $D$ creates a 3D point cloud. In Eq. (3), the incidence field $V$ can measure the observation angle between a 3D plane and the camera. Similar to depthmap $D$ and surface normal map $N$, the incidence field $V$ is invariant to the image cropping and resizing. Consider an image cropping and resizing described as:

$$x' = \Delta K x, \quad \Delta K = \begin{bmatrix} \Delta f_x & 0 & \Delta f_x \\ 0 & \Delta f_y & \Delta f_y \\ 0 & 0 & 1 \end{bmatrix}, \quad K' = \Delta K K,$$

(10)

where $K'$ is the intrinsic after transformation. The surface normal map $N$ and depthmap $D$ after transformation is defined as:

$$N'(x') = N(x) = N(\Delta K^{-1} x'), \quad D'(x') = D(x) = D(\Delta K^{-1} x').$$

(11)

Similarly, the incidence field after transformation is:

$$V'(x') = (K')^{-1} x' = K^{-1} (\Delta K)^{-1} x' = K^{-1} x = v = V(x).$$

(12)

Eq. (12) shows that the incidence field $V$ is a parameterization of the intrinsic matrix that is invariant to image resizing and image cropping. Other invariant parameterizations of the intrinsic matrix, such as the camera field of view (FoV), rely on the central focal point assumption and only cover a 2 DoF intrinsic matrix. An illustration is shown in Fig. 3.

Next, we prove the incidence field $V$ is a monocular 3D prior pixel-wisely determined by monocular depth $D$ and surface normal $N$. Expanding Eq. (4) using the incidence vector $v$ in Eq. (2), we get:

$$n_1 \nabla_x(d) v_x + n_1 \frac{d}{f_x} + n_2 \nabla_y(d) + n_3 \nabla_x = 0.$$  

(13)

Combined with the axis-$y$ constraint, the 2 DoF incidence vector $v = [v_x \quad v_y \quad 1]^T$ is uniquely solved. Eq. (13) gives identical solutions when depthmap $D$ is adjusted by a scalar. This implies that the incidence field $V$ is pixel-wise determined by the up-to-scale depth map and surface normal map, indicating a learnable mapping from the monocular image $I$ to the incidence field $V$ exists.

Given the strong connection between the monocular depthmap $D$ and camera incidence field $V$, we adopt NewCRFs [69], a neural network used in monocular depth estimation, for incidence field estimation. We change the last output head to output a three-dimensional normalized incidence field $\tilde{V}$ with the same resolution as the input image $I$. We adopt a cosine similarity loss defined as:

$$\tilde{V} = D_v(I), \quad L = \frac{1}{N} \sum_{i=1}^{N} \tilde{V}(x_i) \tilde{V}_{gt}(x_i).$$

(14)

We normalize the last dimension of the incidence field to one before feeding to the RANSAC algorithm. That is to say, $V(x_i) = [\tilde{v}_1 \tilde{v}_2 \tilde{v}_3 \quad 1]^T = [v_1 \quad v_2 \quad 1]^T$.

3.4 Monocular Intrinsic Calibration with Incidence Field

Since the network inference executes on GPU device, we adopt a GPU-end RANSAC algorithm to recover the intrinsic $K$ from the incidence field $V$. Unlike a CPU-based RANSAC, we perform fixed $K_e$ iterations of RANSAC without termination. In RANSAC, we use the minimal solver to generate $K_e$ candidates and select the optimal one that maximizes a scoring function (see Fig. 2).

RANSAC w.o Assumption. From Eq. (2), the incidence vector $v$ relates to the intrinsic $K$ as:

$$v = K^{-1} x = \begin{bmatrix} \frac{x - b_x}{f_x} & \frac{y - b_y}{f_y} & 1 \end{bmatrix}^T.$$  

(15)

From Eq. (15), a minimal solver for intrinsic is straightforward. Sample two incidence vectors $v^1 = [v_x^1 \quad v_y^1 \quad 1]^T$ and $v^2 = [v_x^2 \quad v_y^2 \quad 1]^T$. The intrinsic is:

$$\begin{cases} f_x = \frac{x^2 - 2x^2}{v_x^1 - v_x^2} \\ b_x = \frac{1}{2} (x^2 - v_x^1 f_x + x^2 - v_x^2 f_x) \end{cases}, \quad \begin{cases} f_y = \frac{y^2 - 2y^2}{v_y^1 - v_y^2} \\ b_y = \frac{1}{2} (x^2 - v_y^1 f_y + x^2 - v_y^2 f_y) \end{cases}.$$  

(16)

Similarly, the scoring function in $x$-axis and $y$-axis, respectively:

$$\rho_x(f_x, b_x, \{x\}, \{v\}) = \sum_{i=1}^{N_k} \left( \frac{\|x - b_x - v_x^i\| < k_x}{f_x} \right), \quad \rho_y(f_y, b_y, \{x\}, \{v\}) = \sum_{i=1}^{N_k} \left( \frac{\|y - b_y - v_y^i\| < k_y}{f_y} \right).$$  

(17)
Table 2: In-the-Wild Monocular Camera Calibration. We benchmark in-the-wild monocular camera calibration. We use the first 9 datasets for training, and test on all 14 datasets. Except for MegaDepth, we synthesize novel intrinsic by cropping and resizing during training. Note the synthesized images violate the focal point and focal length assumption. [Key: ZS = Zero-Shot, Asm. = Assumptions, Syn. = Synthesized]  

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>NuScenes [40]</td>
<td>Calibrated</td>
<td>Driving</td>
<td>✗</td>
<td>✗</td>
<td>□</td>
<td></td>
<td></td>
</tr>
<tr>
<td>KITTI [32]</td>
<td>Calibrated</td>
<td>Driving</td>
<td>✗</td>
<td>✗</td>
<td>✔</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cityscapes</td>
<td>Calibrated</td>
<td>Driving</td>
<td>✗</td>
<td>✗</td>
<td>✔</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NYUv2 [53]</td>
<td>Calibrated</td>
<td>Indoor</td>
<td>✗</td>
<td>✗</td>
<td>✔</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ARKitScenes</td>
<td>Calibrated</td>
<td>Indoor</td>
<td>✗</td>
<td>✗</td>
<td>✔</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SUN3D [65]</td>
<td>Calibrated</td>
<td>Indoor</td>
<td>✗</td>
<td>✗</td>
<td>✔</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MVImNet [68]</td>
<td>SfM</td>
<td>Object</td>
<td>✗</td>
<td>✗</td>
<td>✔</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Objectron [1]</td>
<td>SfM</td>
<td>Object</td>
<td>✗</td>
<td>✗</td>
<td>✔</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MegaDepth [44]</td>
<td>SfM</td>
<td>Outdoor</td>
<td>✗</td>
<td>✗</td>
<td>✔</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3: Comparisons to Monocular Camera Calibration with Geometry on GSV dataset [4]. We follow the training and testing protocol of [38]. For a fair comparison, we convert the estimated intrinsic to camera FoV on the y-axis direction, following [30, 38], and report our results w/ and w/o the assumptions.  

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>9.47</td>
<td>4.37</td>
<td>3.59</td>
<td>3.07</td>
<td>2.49</td>
<td>2.47</td>
</tr>
<tr>
<td>Median</td>
<td>4.42</td>
<td>3.58</td>
<td>2.72</td>
<td>2.33</td>
<td>1.96</td>
<td>1.92</td>
</tr>
</tbody>
</table>

where \( N_k \) and \( k_x/k_y \) are the number of sampled pixels and the threshold for axis-\(x/y\) directions.

**RANSAC w/ Assumption.** If a simple camera model is assumed, i.e., intrinsic only has an unknown focal length, it only needs to estimate 1-DoF intrinsic. We enumerate the focal length candidates as:

\[
\{ f \} = \{ f_{\text{min}} + \frac{i}{N_f} (f_{\text{max}} - f_{\text{min}}) \mid 0 \leq i \leq N_f \}. \tag{18}
\]

The scoring function under the scenario is defined as the summation over \(x\)-axis and \(y\)-axis:

\[
ρ(f, \{ x \}, \{ y \}) = ρ_x(f_x, w/2, \{ x \}, \{ y \}) + ρ_y(f_y, h/2, \{ x \}, \{ y \}). \tag{19}
\]

### 3.5 Downstream Applications

**Image Crop & Resize Detection and Restoration.** Eq. (10) defines a crop and resize operation:

\[
\mathbf{x}' = \Delta \mathbf{K} \mathbf{x}, \quad \mathbf{K}' = \Delta \mathbf{K} \mathbf{K}^{-1}, \quad \mathbf{I}'(\mathbf{x}') = \mathbf{I}(\Delta \mathbf{K} \mathbf{x}) = \mathbf{I}(\mathbf{x}). \tag{20}
\]

When a modified image \(\mathbf{I}'\) is presented, our algorithm calibrates its intrinsic \(\mathbf{K}'\) and then:

Case 1: The original intrinsic \(\mathbf{K}\) is known. E.g., we obtain \(\mathbf{K}\) from the image-associated EXIF file [3]. Image manipulation is computed as \(\Delta \mathbf{K} = \mathbf{K}' \mathbf{K}^{-1}\). A manipulation is detected if \(\Delta \mathbf{K}\) deviates from an identity matrix. The original image restores as \(\mathbf{I}(\mathbf{x}) = \mathbf{I}'(\Delta \mathbf{K} \mathbf{x})\). Interestingly, the four corners of image \(\mathbf{I}'\) are mapped to a bounding box in original image \(\mathbf{I}\) under manipulation \(\Delta \mathbf{K}\). We thus quantify the restoration by measuring the bounding box. See Fig. 4.

Case 2: The original intrinsic \(\mathbf{K}\) is unknown. We assume the genuine image possess an identical focal length and central focal point. Any resizing and cropping are detected when matrix \(\mathbf{K}'\) breaks this assumption. Note, the rule cannot detect aspect ratio preserving resize or centered crop. We restore the original image by defining an inverse operation \(\Delta \mathbf{K}\) restore \(\mathbf{K}'\) to an intrinsic fits the assumption.

**3D Sensing Related Tasks.** Intrinsic estimation can enable multiple applications. E.g., depthmap to point cloud, uncalibrated two-view pose estimation, etc.

### 4 Experiments

**Implementation Details** Our network is trained using the Adam optimizer [32] with a batch size of 8. The learning rate is \(1e^{-5}\), and the training process runs for 20,000 steps. Our RANSAC algorithm
We then uniformly random resize up to two times its size and subsequently crop to a resolution of $640 \times 480$. This brings three benefits. First, the assumption restricts their training to panorama images. In contrast, our model is trainable with any calibrated images. This yields improved generalization, as shown in Table 2.

### Evaluation Metrics

We follow [29] in assessing camera intrinsic estimation using the metrics:

$$
e_f = \frac{\|f_x^\prime - f_x\|}{f_x}, \quad e_y = \frac{\|f_y^\prime - f_y\|}{f_y}, \quad e_{bx} = 2 \cdot \frac{\|b_x^\prime - b_x\|}{w}, \quad e_{by} = 2 \cdot \frac{\|b_y^\prime - b_y\|}{h}.
$$

(21)

We additionally consolidate the performance along the x-axis and y-axis into comprehensive measurements, $e_f = \max(e_{fx}, e_{fy})$ and $e_b = \max(e_{bx}, e_{by})$.

### 4.1 Monocular Camera Calibration In-The-Wild

**Datasets.** Our method is trained whenever a calibrated intrinsic is provided, making it applicable to a wide range of publicly available datasets. In Tab. 2, we incorporate datasets of different application scenarios, including indoor, outdoor scenes, driving, and object-centric scenes. Many of the datasets utilize only a single type of camera for data collection, resulting in a scarcity of intrinsic variations. Similar to [38], we employ random resizing and cropping to synthesize more intrinsic variations. We report with the former two variations: one assumes a central focal point, and another does not. We report with the latter whenever the input image fits the assumption. From Tab. 2, our method demonstrates superior generalization across multiple unseen datasets. Further, the result w/ assumption outperforms w/o assumption whenever the input images fit the assumption. Tested on an RTX-2080 Ti GPU, the combined network inference and calibration algorithm runs on average in 87 ms per image.

### 4.2 Monocular Camera Calibration with Geometry

Methods in this line of research hold a Manhattan World assumption, positing that images consist of planes that are either parallel or perpendicular to each other. Stated in Tab. 1, baselines [28, 30, 38] relax the assumption to training data. Our method imposes no other assumption except for undistorted images in both training and testing.

This brings three benefits. First, the assumption restricts their training to panorama images. In contrast, our model is trainable with any calibrated images. This yields improved generalization,
Figure 4: Image Crop & Resize Detection and Restoration. Image editing, including cropping and resizing changes intrinsic. As in Sec. 3.5, monocular calibration is applicable to detect and restore image manipulations. We visualize the zero-shot samples on ScanNet and Waymo. More examples are in Supp.

4.3 Monocular Camera Calibration with Objects

We compare to the recent object-based camera calibration method FaceCalib [29]. The baseline employs a face alignment model to calibrate the intrinsic over a video. Both [29] and ours perform zero-shot prediction. We report performance using Tab. 2 model. Compared to [29], our method is more general as it does not assume a human face present in the image. Meanwhile, [29] calibrates over a video, while ours is a monocular method. For a fair comparison, we report the video-based results as an averaged error over the videos. We report results w/ and w/o assuming a simple camera model. Since the tested image has a central focal point, when the assumption applied, the error of the focal point diminished. In Tab. 4, we outperform SoTA substantially.

4.4 Downstream Applications

Image Crop & Resize Detection and Restoration. In Sec. 3.5, our method also addresses the detection and restoration of geometric manipulations in images. Using the model reported in Tab. 2, we benchmark its performance in Tab. 5. Random manipulations following Sec. 4.1 contribute to 50% of both train and test sets, and the other 50% are genuine images. In Tab. 5, we evaluate restoration with mIOU and report detection accuracy (i.e. binary classification of genuine vs edited images). From the table, our method generalizes to the unseen dataset, achieving an averaged mIOU of 0.680. Meanwhile, we substantially outperform the baseline, which is trained to directly regress the intrinsic. The improvement suggests the benefit of the incidence field as an invariant intrinsic parameterization. Beyond performance, our algorithm is interpretable. In Fig. 4, the perceived image geometry is interpretable for humans.
Uncalibrated Two-View Camera Pose Estimation. With correspondence between two images, one can infer the fundamental matrix [25]. However, the pose between two uncalibrated images is determined by a projective ambiguity. Our method eliminates the ambiguity with monocular camera calibration. In Tab. 6, we benchmark the uncalibrated two-view pose estimation and compare it to recent baselines. The result is reported using the model benchmarked in Tab. 2 and assumes unique intrinsic for both images. We perform zero-shot in ScanNet. For MegaDepth, it includes images collected over the Internet with diverse intrinsics. Interestingly, in ScanNet, our uncalibrated method outperforms a calibrated one [44]. In Supp, we plot the curve between pose performance and intrinsic quality. The challenging setting suggests itself an ideal task to evaluate the intrinsic quality.

In-the-Wild Monocular 3D Sensing. In-the-wild 3D sensing, e.g., monocular 3D object detection [9], requires intrinsic to connect the 3D and 2D space. In Fig. 5, despite 3D bounding boxes being accurately estimated, incorrect intrinsic produces suboptimal 2D projection. Our calibration results suggest monocular incidence field estimation is another essential monocular 3D sensing task.

5 Conclusion

We calibrate monocular images through a novel monocular 3D prior referred as incidence field. The incidence field is a pixel-wise parameterization of intrinsic invariant to image resizing and cropping. A RANSAC algorithm is developed to recover intrinsic from the incidence field. We extensively benchmark our algorithm and demonstrate robust in-the-wild performance. Beyond calibration, we show multiple downstream applications that benefit from our method.

Limitation. In real application, whether to apply the assumption still requires human input.

Broader Impacts. Our method detects image spacial editing. If the Exif files are provided, it further localizes the modification region. Shown in Fig. 6, it helps improve press image authenticity.
References

[14] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In CVPR, 2016. 7
[18] Bibhash Das, Mrutyunjay Biswal, Abhranta Panigrahi, and Manish Okade. CNN based image resizing and resize factor classification for forensic applications. In ICORT, 2021. 3
[23] Alexander Grabner, Peter Roth, and Vincent Lepetit. GP3C: Geometric projection parameter consensus for joint 3D pose and focal length estimation in the wild. In ICCV, 2019. 1, 2, 3


[60] Zachary Teed and Jia Deng. DeepV2D: Video to depth with differentiable structure from motion. In ICLR, 2020. 3

[61] Rafael Von Gioi, Jeremie Jakubowicz, Jean-Michel Morel, and Gregory Randall. LSD: A fast line segment detector with a false detection control. TPAMI, 2008. 3


[64] Scott Workman, Menghua Zhai, and Nathan Jacobs. Horizon lines in the wil. In BMVC, 2016. 1, 3


[68] Xianggang Yu, Mutian Xu, Yidan Zhang, Haolin Liu, Chongjie Ye, Yushuang Wu, Zizheng Yan, Chenming Zhu, Zhangyang Xiong, Tianyou Liang, Guanying Chen, Shuguang Cui, and Xiaoguang Han. MVImgNet: A large-scale dataset of multi-view images. In CVPR, 2023. 7


