DiffCalib: Reformulating Monocular Camera Calibration as Diffusion-Based Dense Incident Map Generation

Xiankang He[1,2*], Guangkai Xu[3*], Bo Zhang[3], Hao Chen[3], Ying Cui[1,2], Dongyan Guo[1,2†]

[1]College of Computer Science and Technology, Zhejiang University of Technology
[2]Zhejiang Key Laboratory of Visual Information Intelligent Processing
[3]State Key Lab of CAD & CG, Zhejiang University
{hexiankang577, guangkai.xu, zhangboknight, stanzju}@gmail.com,
{cuiying, guodongyan}@zjut.edu.cn,

Abstract

Monocular camera calibration is a key precondition for numerous 3D vision applications. Despite considerable advancements, existing methods often hinge on specific assumptions and struggle to generalize across varied real-world scenarios, and the performance is limited by insufficient training data. Recently, diffusion models trained on expansive datasets have been confirmed to maintain the capability to generate diverse, high-quality images. This success suggests a strong potential of the models to effectively understand varied visual information. In this work, we leverage the comprehensive visual knowledge embedded in pre-trained diffusion models to enable more robust and accurate monocular camera intrinsic estimation. Specifically, we reformulate the problem of estimating the four degrees of freedom (4-DoF) of camera intrinsic parameters as a dense incident map generation task. The map details the angle of incidence for each pixel in the RGB image, and its format aligns well with the paradigm of diffusion models. The camera intrinsic can then be derived from the incident map with a simple non-learning RANSAC algorithm during inference. Moreover, to further enhance the performance, we jointly estimate a depth map to provide extra geometric information for the incident map estimation. Extensive experiments on multiple testing datasets demonstrate that our model achieves state-of-the-art performance, gaining up to a 40% reduction in prediction errors. Besides, the experiments also show that the precise camera intrinsic and depth maps estimated by our pipeline can greatly benefit practical applications such as 3D reconstruction from a single in-the-wild image.

Introduction

Monocular camera calibration (Zhu et al. 2023; Jin et al. 2022; Hold-Geoffroy et al. 2018; Lee et al. 2020) aims to estimate the intrinsic properties of a camera from a single image, which is important for many downstream tasks of 3D scene reconstruction and understanding, as well as other visual applications (Zhang et al. 2020a,b). Existing methods often rely on geometric principles like Manhattan world assumption (Coughlan and Yuille 1999), or are based on specific objects like checkerboards or human faces (Hu et al. 2023), and therefore can hardly generalize to diverse real scenarios. To alleviate the reliance mentioned above, Zhu (Zhu et al. 2023) introduced an incident field concept, which defines the direction between the 3D point cloud and the camera's optical center, allowing intrinsic recovery using a simple RANSAC algorithm (Fischler and Bolles 1981). While this approach shows promising generalization, it struggles with accurate and confident estimates due to limited training data like the ill-pose property of monocular depth estimation.

Recently, a series of advanced approaches (Ke et al. 2023; Xu et al. 2024; Fu et al. 2024) for monocular depth estimation have emerged, which leverage the robust knowledge priors embedded within Stable Diffusion models. Through strategic fine-tuning protocols, these methods have demonstrated exceptional capabilities in achieving commendable zero-shot generalization.

In this paper, motivated by the successful visual knowledge transfer from image generation to depth estimation, as seen in Marigold, we propose to solve the monocular camera calibration problem by reformulating it as an incident map generation task. The incident map details the angle of incidence for each pi‘xel in the RGB image and its format aligns well with the paradigm of diffusion models. Thus, it can be learned by properly fine-tuning and enforcing the Stable Diffusion models. With such design, we can effectively address the community's focus on the robustness and accuracy by leveraging the advantages of diffusion models. Specifically, we freeze the VAE encoder and decoder of Stable Diffusion, and fine-tune the U-Net to learn the noise added to the incident map. It can bring two main advantages: 1) We prove that the rich visual information of Stable Diffusion models can benefit not only perception tasks but also the camera characteristics estimation. 2) By regarding the estimation process as a probabilistic one, the confidence of the predicted incident map will be increased by de-noising from different noise maps and ensembling the corresponding results.

Additionally, Yin et al. (Yin et al. 2021) analyzed the connection between camera intrinsic parameters and depth maps in single-image 3D reconstruction, highlighting their inherent relationship. To enhance performance, we incorporate

*These authors contributed equally.
†Corresponding author.

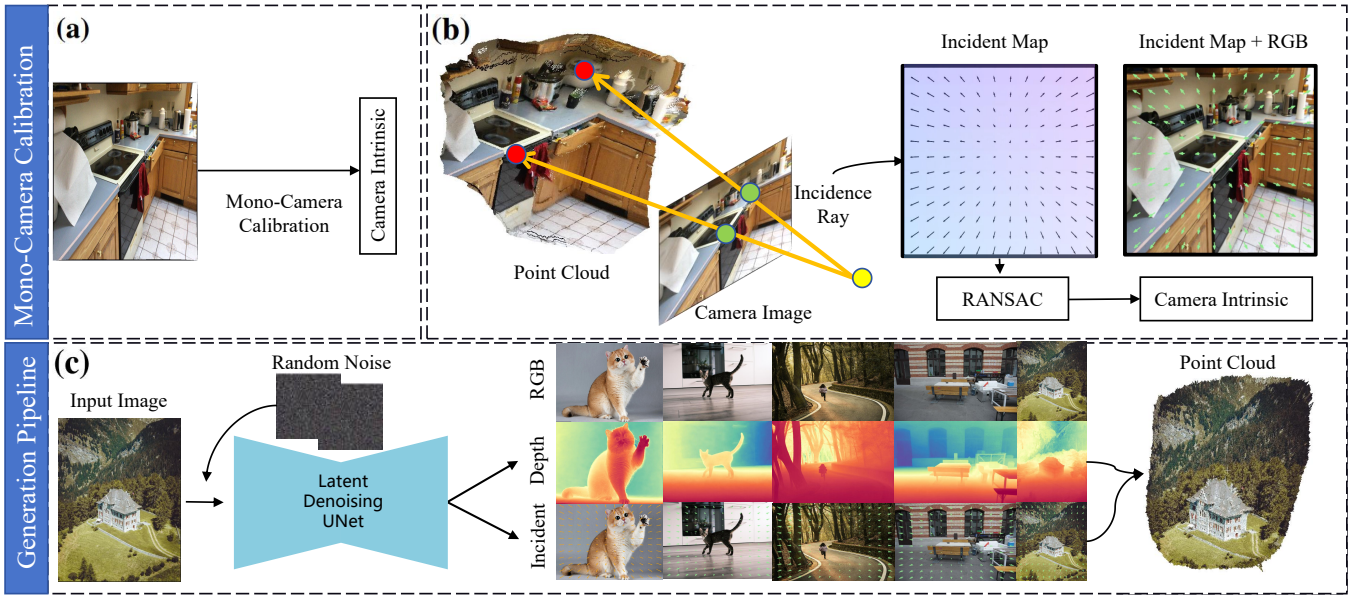

Figure 1: We reformulate monocular camera calibration as a diffusion-based incident map generation task. (a) Our pipeline enables robust camera intrinsic estimation from a single image. (b) The 'incidence map' represents incident rays pointing from the camera image pixels to the 3D point cloud. The camera intrinsic can be derived from the incident map with the RANSAC algorithm. (c) Our pipeline leverages the paradigm of latent diffusion models, takes the RGB image and random Gaussian noise as input, and generates the incidence map and depth map together. Subsequently, in-the-wild 3D reconstruction is enabled with the predicted depth and intrinsic.

depth information into our pipeline, forcing the network to concurrently estimate both the incident map and the depth map from an input RGB image. Our experiments demonstrate that this approach improves results in both modalities.

Using the estimated incident map, the camera's intrinsic parameters can be accurately derived through the non-learning-based RANSAC (Fischler and Bolles 1981) algorithm. Extensive quantitative and qualitative experiments show that our method outperforms recent approaches and achieves state-of-the-art results. Comprehensive ablation studies verify the effectiveness of each component. Additionally, leveraging the estimated depth map allows us to project the 2D image into 3D space, facilitating 3D reconstruction from a single in-the-wild image.

Overall, our contributions can be summarized as follows.

- To the best of our knowledge, this is the first work to leverage the visual knowledge priors of diffusion models to reformulate the camera intrinsic estimation as the task of generating a dense incident map. This approach significantly improves the robustness and accuracy of the estimates.
- We introduce a method to jointly estimate the incident map and the depth map, leveraging their intrinsic relationships to enhance the performance of both.
- Utilizing the predicted depth map and camera intrinsics derived from the incident map, our approach can benefit downstream applications such as 3D reconstruction from a single image, even in challenging in-the-wild scenes.

Related Work

Monocular Camera Calibration

Monocular camera calibration is vital in computer vision, focusing on geometric considerations and object properties. Early methods used geometric principles like the Manhattan world assumption (Coughlan and Yuille 1999) and objects such as chessboards (Zhang 2000) or line segments (Von Gioi et al. 2008; Akinlar and Topal 2011). While these methods offered diverse approaches, they depended on specific assumptions, limiting their generalization. More recent methods used real-world objects like faces (Hu et al. 2023) or other items (Grabner, Roth, and Lepetit 2019; Chen, Chin, and Li 2019; Sturm 2005), but still required specific objects, restricting their application in varied scenarios. In contrast, our method reformulates calibration as an image generation task. By leveraging the diffusion model, our method predicts the four degrees of freedom (4-DoF) of camera intrinsic parameters using a single undistorted image, eliminating the need for conventional assumptions and specific objects.

Learnable Monocular 3D Priors

Zhu et al.(Zhu et al. 2023) introduced the incidence field, defined as the incidence rays between 3D points and 2D pixels, to estimate camera intrinsics by exploiting monocular 3D priors. The most well-known 3D priors in computer vision are monocular depth(Yin et al. 2021; Ranftl, Bochkovskiy, and Koltun 2021; Ranftl et al. 2022; Yin et al. 2023; Xu et al. 2022, 2024) and surface normals (Bae, Budvytis, and Cipolla 2021; Xu et al. 2024), which are fundamental for

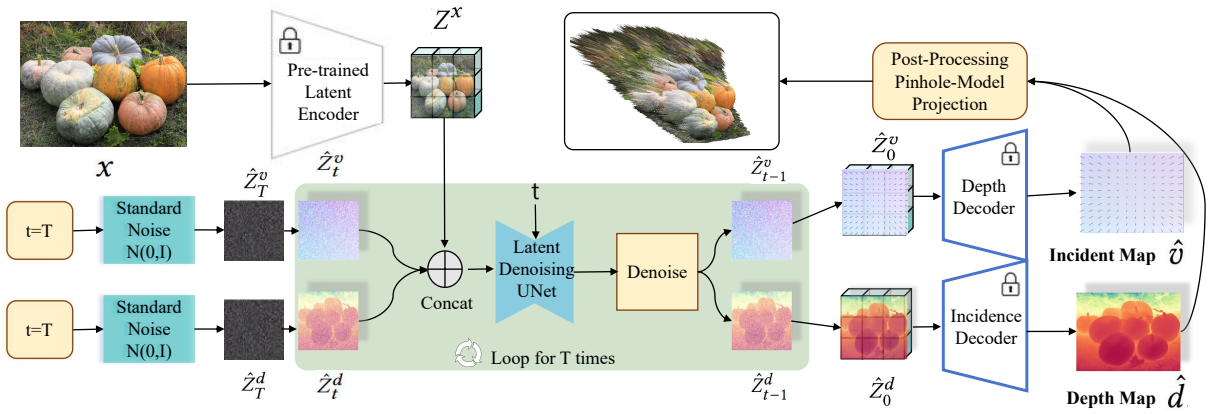

Figure 2: Overview of the generation pipeline. Given an image x, we generate the incident map $\hat{v}$ and depth map $\hat{d}$ using the denoising U-Net from two randomly sampled gaussian noises $\hat{Z}_T^v$ and $\hat{Z}_T^d$. The generated $\hat{v}$ and $\hat{d}$ are projected into 3D space to recover the 3D scene shape. It is worth mentioning that the denoising process in the green part loops for T times. Please note that the Incidence Decoder and Depth Decoder utilize the same freeze decoder.

transitioning from 2D to 3D tasks by providing essential information about scene depth and surface orientation. Recently, Jin et al.(Jin et al. 2023) introduced the perspective field for single image camera calibration, offering per-pixel information about the camera view through an up vector and latitude value, effectively capturing local perspective properties. Unlike the perspective field approach requiring panorama images for training, the incidence field only needs undistorted images, providing a valuable solution for in-the-wild monocular camera calibration. A Transformer-based fully-connected CRFs neural network(Yuan et al. 2022) is used for incidence field estimation. However, the incidence field approach is limited by insufficient training data. Our method addresses this by formulating incidence field estimation as an incidence map generation task, leveraging the robust priors in generative Stable Diffusion models to achieve more generalized and robust in-the-wild single image camera calibration.

Diffusion Models

The Diffusion Denoising Probability Model (DDPM), also known as the Diffusion Model (Ho, Jain, and Abbeel 2020), presents a novel generative approach distinct from GANs (Goodfellow et al. 2014). Renowned for its high-quality generation and controllable synthesis, DDPM has gained increasing popularity across diverse domains. Recently, large language models have also been explored for image generation (Liu et al. 2025a; He et al. 2025) and other tasks (Liu et al. 2025b). The DDPM model trains a denoising encoder to reverse the Markov diffusion process (Song, Meng, and Ermon 2020). Advancements like the latent diffusion model (Rombach et al. 2022) reduce computational costs while preserving high fidelity. Our method leverages the diffusion model's robust priors to generate high-precision incident and depth maps, improving intrinsic parameter estimation and enabling more accurate 3D reconstruction.

Method

The overall pipeline of DiffCalib is shown in Figure 2. Our approach further formulates the incidence field estimation as an incidence map generation task, with rich visual knowledge of the pre-trained diffusion models leveraged to enhance the robustness of calibration. Additionally, the simultaneous estimation of depth maps and incident maps facilitates performance boosting and applications like 3D reconstruction from a single image. To begin with, let's introduce the concepts of the incident map, and reformulate it as a diffusion-based generation.

Incident Map

The incident map contains the collection of incident rays originating from points within the scene and passing through corresponding pixels on the camera's imaging plane. It outlines the array of rays extending from each pixel position on the imaging plane to the camera's focal point.

Take the pinhole camera model as an example. Mathematically, for any pixel p of coordinate (x, y) on the imaging plane, the incident map vector v, denoted as $V(p)$, can be expressed as follows.

$$v = V(p) = V(x, y) = \begin{pmatrix} \frac{x-b_x}{f_x} & \frac{y-b_y}{f_y} & 1 \end{pmatrix}^\top \quad (1)$$

Where b_x and b_y denote the optical center location along the x-axis and y-axis, and f_x and f_y represent the focal length along the x-axis and y-axis, respectively.

The incident map emerges as a crucial 3D prior due to its ability to convey essential information regarding the camera's viewing perspective. In contrast to directly estimating the 4-DoF camera intrinsic parameters, the dense incident map bears a closer resemblance to natural images and is invariant to image transformations such as cropping and resizing. This similarity enables it to potentially leverage the extensive knowledge priors of networks that have been pre-trained on a diverse array of real-world scenes.

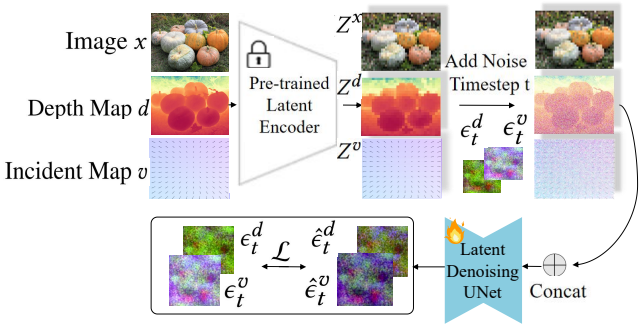

Figure 3: Overview of the training pipeline. We freeze the latent encoder and encode the input image $\mathbf{x}$, incident map $\mathbf{v}$, and depth map $\mathbf{d}$ into the latent space. Then, the U-Net is trained to predict the noise added to the depth and incident map latent codes, denoted as $\hat{\epsilon}_t^d$ and $\hat{\epsilon}_t^v$, respectively. The loss function is computed between the estimated noise and the added ground-truth noise.

Diffusion-Based Generation

With the dense representation of the incident map, we reformulate the monocular camera calibration task as an incident map generation process by leveraging the rich knowledge priors of the Stable Diffusion v2.1 model. The Stable Diffusion is composed of a VAE autoencoder that transfers images to the latent space, and a U-Net that estimates the noise added to the image.

During training, the incident map is firstly encoded to the latent space as Z^v with the pre-trained VAE encoder. Then, we gradually add random sample Gaussian noise to get the noising incident latent codes Z_t^v:

$$Z_t^v = \sqrt{\bar{\alpha}_t} Z_0^v + \sqrt{1-\bar{\alpha}_t}\epsilon \tag{2}$$

where $Z_0^v = Z^v$ is the initial incident map and $\bar{\alpha}_t$ is noise scheduler that controls sample quality. The timestep $t \sim (1, T)$ and $\epsilon \sim \mathbf{N}(0, I)$. Then, the noisy encoded incident map passes through the U-Net ϵ_θ to predict the estimated noise $\hat{\epsilon}$. The model is trained by minimizing the L_2 loss of the estimated noise $\hat{\epsilon}$ and the ground-truth added noise ϵ

During inference, we denoise the input Z_t^v to Z_{t-1}^v step by step from Z_T^v to Z_0^v with the trained denoiser U-Net. Finally, the estimated incident map is recovered from Z_0^v with the pre-trained VAE decoder.

DiffCalib

Our approach leverages pre-training on Stable Diffusion v2.1 (Rombach et al. 2022), aiming to reformulate incidence field estimation as an incidence map generation task. This allows us to use the robust priors from generative Stable Diffusion models for more generalized and robust in-the-wild single-image camera calibration. However, a challenge arises due to the uniform structural distribution of the incident map, which only comprises four parameters, differing significantly from the pre-trained image data. The similarity between pixels and their values makes it difficult for the diffusion model to effectively utilize image generation priors. Bridging this gap is crucial for the model's effectiveness.

Enhance Incident Map Generation by Jointly Learning with Depth Map. Inspired by Yin et al. (Yin et al. 2021), we recognize the association between depth information and camera intrinsic parameters, highlighting the correlation between incident maps and depth maps. Using depth maps, we can map incidence rays to latent 3D space, and incident maps can project depth maps into 3D space. Thus, we jointly incorporate depth and incident maps into our model to enhance incident map generation.

Our training pipeline is illustrated in Figure 3. When inputting the image x and incident map v, the paired depth map d is included, replicated into three channels to resemble RGB. The frozen VAE Encoder $E(\cdot)$ encodes x, d, and v as $Z^x = E(x)$, $Z^d = E(d)$, and $Z^v = E(v)$. Z^d and Z^v are added with multi-resolution noises (Kasiopy 2023) ϵ_t^d and ϵ_t^v to form Z_t^d and Z_t^v. These are concatenated as (Z^x, Z_t^v, Z_t^d).

We triple the input channel of the U-Net in Stable Diffusion v2.1 to accommodate 12 channels. The joint latent codes Z_t^v and Z_t^d form a 3D representation Z_t^P, allowing the denoiser to utilize the joint latent variable representation for denoising. The denoise function is:

$$\hat{\epsilon}_t^P = \epsilon_\theta(Z^x, Z_t^P, t) \tag{3}$$

where $\hat{\epsilon}_t^P$ is split into $\hat{\epsilon}_t^v$ and $\hat{\epsilon}_t^d$. The loss is:

$$\begin{aligned}\mathcal{L} = \mathbb{E}_{x,v,d,\epsilon_t^v \sim \mathcal{N}(0,I), t\sim U(T)} \|\epsilon_t^v - \hat{\epsilon}_t^v\|_2^2 + \\ \mathbb{E}_{x,v,d,\epsilon_t^d \sim \mathcal{N}(0,I), t\sim U(T)} \|\epsilon_t^d - \hat{\epsilon}_t^d\|_2^2\end{aligned} \tag{4}$$

Inference Performance Improvement with the Ensemble Process. During the inference phase, we use a frozen VAE encoder to convert the input image x into latent code Z^x. We then generate ensemble-size noises $\hat{Z}_T^v$ and $\hat{Z}_T^d$ for the incident map and depth, respectively. Each noise sample is paired with the image to ensure diversity.

To enhance accuracy, we average the generated incident noise, where less accurate pixels are refined through multiple generations while accurate pixels remain stable. This approach improves the reliability of the incident maps. The combined latent codes $(Z^x, \hat{Z}_T^v, \hat{Z}_T^d)$ are then processed through U-Net for multistep denoising. The output latent codes $\hat{Z}_0^v$ and $\hat{Z}_0^d$ are decoded by the frozen VAE decoder to produce the ensemble incident maps $\hat{v}$ and depth maps $\hat{d}$. The ensemble maps $\hat{v}$ and $\hat{d}$ are averaged to obtain the final incident map and depth map. While the incident map is directly reconstructed, the depth map, initially in three channels, is condensed into a single channel through averaging.

Monocular Intrinsic Calibration from Incident Map. With the reconstructed incident map $\hat{v}$, we use a RANSAC method without assumptions to recover the camera's intrinsic matrix K. from the incident map. By leveraging the relationship between the incidence vector v and the camera intrinsic parameters, the intrinsic matrix can be directly inferred.

With the 2D pixel location of the image $\mathbf{p} = [x,\ y,\ 1]^T$ and camera intrinsic K:

$$\mathbf{K} = \begin{bmatrix} f_x & 0 & b_x \\ 0 & f_y & b_y \\ 0 & 0 & 1 \end{bmatrix} \tag{5}$$

, we can randomly sample two incidence vectors:

$$\mathbf{v}_1 = \mathbf{K}^{-1} \begin{pmatrix} x_1 & y_1 & 1 \end{pmatrix}^\top = \begin{pmatrix} \frac{x_1 - b_x}{f_x} & \frac{y_1 - b_y}{f_y} & 1 \end{pmatrix}^\top, \quad (6)$$

$$\mathbf{v}_2 = \mathbf{K}^{-1} \begin{pmatrix} x_2 & y_2 & 1 \end{pmatrix}^\top = \begin{pmatrix} \frac{x_2 - b_x}{f_x} & \frac{y_2 - b_y}{f_y} & 1 \end{pmatrix}^\top \quad (7)$$

Then the camera's intrinsic matrix is estimated using the RANSAC algorithm and a minimal solver follows WildCamera (Zhu et al. 2023). Based on the relationship between the incidence vector and the camera's intrinsic parameters, we can directly derive the focal lengths as well as the pixel coordinates of the optical center:

$$f_x = \frac{x_1 - x_2}{v_x^1 - v_x^2}, \quad b_x = \frac{1}{2}(x_1 - v_x^1 f_x + x_2 - v_x^2 f_x)$$
$$f_y = \frac{y_1 - y_2}{v_y^1 - v_y^2} \quad b_y = \frac{1}{2}(y_1 - v_y^1 f_y + y_2 - v_y^2 f_y)$$

If we assume that the optical center is positioned at the center of the image, and the camera model adheres to the pinhole model, the estimation of the camera's intrinsic parameters can be simplified to a 1-Degree of Freedom (1-DoF) task. Specifically, we can estimate the focal length of the camera by enumerating candidate values.

Downstream Application: 3D Reconstruction

With the generated depth map $\hat{d}$ and estimated camera intrinsic f, we can reconstruct the 3D point cloud using a simple pinhole camera model. However, the depth map we generate is the affine-invariant depth, which includes a concept of shift that leads to the distortion of the reconstructed point cloud. To mitigate this issue, we leverage the frozen shift model (Yin et al. 2021) to recover the shift. With the recovery of the shift in our depth map $\hat{d}$, we can obtain the point cloud $\mathbf{P}$ as:

$$\mathbf{P} = \hat{d} \cdot \begin{pmatrix} \frac{x - b_x}{f_x} & \frac{y - b_y}{f_y} & 1 \end{pmatrix}^\top \quad (8)$$

Here, x and y represent the pixel coordinates of the depth map, while the estimated $f = f_x = f_y$ and b_x, b_y denote the location of the map's center in our pinhole camera model.

Experiments

Dataset and Evaluation Protocol

Training Datasets We choose Hypersim (Roberts et al. 2021) as our primary training dataset for incident map and depth map generation. This dataset comprises 461 synthetic indoor scenes with depth information and consistent ground-truth camera intrinsic parameters of $[889, 889, 512, 384]$ across all scenes. We use 365 scenes for training, following the recommended setup. To increase the variety of training scenarios, we incorporate additional datasets: NuScenes (Caesar et al. 2020), KITTI (Geiger et al. 2013), CityScapes (Cordts et al. 2016), NYUv2 (Silberman et al. 2012), SUN3D (Xiao, Owens, and Torralba 2013), ARKitScenes (Baruch et al. 2021), Objectron (Ahmadyan et al. 2021), and MVImgNet (Yu et al. 2023). These datasets are used with their camera intrinsic parameters but without depth information. For consistency, we replace the depth input with a copied image input to match the network input. To introduce variations in intrinsic parameters, we augment the intrinsic settings by randomly enlarging images up to twice their size and then cropping them to a suitable size, following the approach in (Lee et al. 2021). This augmentation addresses the scarcity of intrinsic variations within the dataset, ensuring a robust training process.

Testing Datasets For monocular camera calibration, our evaluation encompasses datasets such as Waymo (Sun et al. 2020), RGBD (Sturm et al. 2012), ScanNet (Dai et al. 2017), MVS (Fuhrmann, Langguth, and Goesele 2014), and Scenes11 (Chang et al. 2015). We ensure alignment with the benchmark provided by WildCamera (Zhu et al. 2023) for this task.

Evaluation Protocol For camera intrinsic estimation assessment, we adhere to the prescribed evaluation protocol in (Zhu et al. 2023), employing the metrics:

$$e_f = \max\left(\frac{|f_x' - f_x|}{f_x}, \frac{|f_y' - f_y|}{f_y}\right),$$
$$e_b = \max\left(2 \cdot \frac{|b_x' - b_x|}{w}, 2 \cdot \frac{|b_y' - b_y|}{h}\right) \quad (9)$$

where f_x and f_y represent the focal lengths along the two axes, b_x and b_y denote the location of the optical center, w and h represent the width and height of the image, respectively.

Implementation Details

We leverage the pre-training model provided by Stable Diffusion v2.1 (Rombach et al. 2022), wherein we freeze the VAE encoder and decoder, focusing solely on training the U-Net. This training regimen adheres to the original pre-training setup with a v-objective. Moreover, we configure the noise scheduler of DDPM with 1000 steps to optimize the training process. The training regimen comprises 30,000 iterations, with a batch size of 16. To accommodate the training within a single GPU, we accumulate gradients over 16 steps. We employ the Adam optimizer with a learning rate of 3×10^{-5}. Typically, achieving convergence during our training process necessitates approximately 12 hours when executed on a single Nvidia RTX A800 GPU card. We set the ensemble size as 10, meaning we aggregate predictions from 10 inference runs for each image.

Quantitative Comparison

Quantitative Comparison of Monocular Camera Calibration. We present the monocular camera calibration results separately for both seen and unseen datasets in Table 1 and Table 2, respectively.

In Table 1, for a fair comparison, we utilize the same data as WildCamera (Zhu et al. 2023) to train our method specifically for the incident map and evaluate the metrics on the test split of the seen dataset. Our approach achieves significant improvements on most datasets. Furthermore, in Table 2, our method's evaluation on zero-shot in-the-wild datasets demonstrates superior generalization in diverse

Methods	NuScenes		KITTI		CitySpace		NYUv2		SUN3D		ARKitScenes		MVImgNet	
	$e_f \downarrow$	$e_b \downarrow$	$e_f \downarrow$	$e_b \downarrow$	$e_f \downarrow$	$e_b \downarrow$	$e_f \downarrow$	$e_b \downarrow$	$e_f \downarrow$	$e_b \downarrow$	$e_f \downarrow$	$e_b \downarrow$	$e_f \downarrow$	$e_b \downarrow$
Perspective CVPR'23	0.378	0.286	0.631	0.279	0.624	0.316	0.261	0.348	0.325	0.367	0.260	0.385	0.601	0.311
WildCamera NeurIPS'23	0.102	0.087	0.111	0.078	0.108	0.110	0.086	0.174	0.113	0.205	0.140	0.243	0.101	0.081
DiffCalib (w/o depth)	0.075	**0.022**	0.087	0.094	0.062	0.047	0.057	**0.022**	0.059	**0.023**	0.107	**0.027**	0.108	**0.031**
DiffCalib	**0.026**	0.039	**0.021**	**0.074**	**0.052**	**0.045**	**0.013**	0.040	**0.028**	0.043	**0.069**	0.048	**0.078**	0.074

Table 1: Monocular camera calibration was conducted on the testing split of trained datasets. We followed the benchmark of WildCamera (Zhu et al. 2023). 'DiffCalib (w/o depth)' represents the pipeline that only uses the incident map. 'DiffCalib' represents the pipeline that jointly utilizes the incident map and the depth map.

Methods	Asm	Waymo		RGBD		ScanNet		MVS		Scenes11	
		$e_f \downarrow$	$e_b \downarrow$	$e_f \downarrow$	$e_b \downarrow$	$e_f \downarrow$	$e_b \downarrow$	$e_f \downarrow$	$e_b \downarrow$	$e_f \downarrow$	$e_b \downarrow$
WildCamera (Zhu et al. 2023)	×	0.210	**0.053**	0.097	0.039	0.128	0.041	0.170	**0.028**	0.170	0.044
DiffCalib (w/o depth)	×	0.188	**0.053**	0.092	**0.018**	0.089	0.041	0.135	0.032	**0.108**	**0.029**
DiffCalib	×	**0.145**	**0.053**	**0.084**	0.040	**0.055**	**0.036**	**0.108**	0.036	0.176	0.038
Perspective (Jin et al. 2023)	✓	0.444	0.020	0.166	0.000	0.189	0.010	0.185	0.000	0.211	0.000
WildCamera	✓	0.157	0.020	0.067	0.000	0.109	0.010	0.127	0.000	0.117	0.000
DiffCalib (w/o depth)	✓	0.246	0.020	**0.052**	0.000	0.071	0.010	0.112	0.000	**0.081**	0.000
DiffCalib	✓	**0.120**	0.020	0.062	0.000	**0.042**	0.010	**0.081**	0.000	0.146	0.000

Table 2: Monocular camera calibration on zero-shot datasets was conducted following the benchmark established by WildCamera (Zhu et al. 2023). **Asm** represents the assumption that the image center of the simple camera model is fixed as the optical center, resulting in **consistent outcomes**. Our approach achieves state-of-the-art performance on zero-shot datasets.

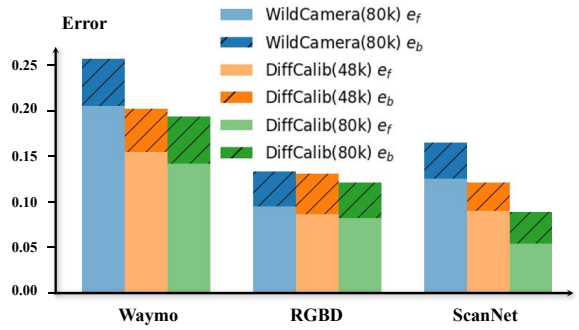

Figure 4: Comparison with limited data: We compared the model training using less data to achieve greater results, while the inclusion of additional data leads to even better performance.

scenes, outperforming others in real-world scenarios. Comparing DiffCalib with DiffCalib (w/o depth) reveals robust joint-learning capabilities, improving most datasets except RGBD (Sturm et al. 2012) and Scenes11 (Chang et al. 2015). The RGBD discrepancy may be due to uncalibrated default intrinsic parameters (Zhu et al. 2023), and the Scenes11 dataset's complexity arises from randomly shaped and moving objects. Thus, using only incident maps without depth estimation performs better on these two datasets.

Quantitative Comparison of Zero-Shot 3D Scene Reconstruction To demonstrate the robustness and accuracy of our reconstruction method, we compare it against LeReS (Yin et al. 2021), which also uses a single image for 3D scene reconstruction. The results, shown in Table 3, utilize Chamfer l_1 distance (C-l_1) and F-score with a 5cm threshold. Both methods align predicted point clouds with

Methods	NYU		ScanNet	
	C-$l_1 \downarrow$	F-$score \uparrow$	C-$l_1 \downarrow$	F-$score \uparrow$
LeReS	0.145	0.333	**0.126**	0.377
DiffCalib	**0.127**	**0.433**	0.150	**0.508**

Table 3: Qualitative comparison of 3D reconstruction from single image. We compare the reconstruction performance with LeReS on two zero-shot datasets. The C-l_1 and F-$score$ metrics with a threshold of 5cm are evaluated here.

the ground truth depth, using the identity matrix for pose. Our method shows a 30% improvement in the F-score metric over LeReS. For focal length predictions on ScanNet (Dai et al. 2017), LeReS predicts 980, our method predicts closer to the ground truth of 1165.72, resulting in better performance. Occasional poor C-l_1 results may be due to outlier point clouds.

Ablation Study

Pre-train of U-Net. Table 4 illustrates the performance of camera intrinsic estimation with and without U-Net pre-training. To ensure adequate fitting, we train the network for 30k iterations (10k more than with pre-training) and test it on validation data. The results show significant performance improvement with U-Net pre-training, highlighting the importance of using a generalizable and strong prior in this domain. We present the worst results from two datasets, which largely represent the impact of U-Net pre-training.

Comparison with Less Data. As shown in the Figure 4, with less than 48k of 80k data(about half), DiffCalib still outperforms WildCamera, which demonstrates the superiority of fine-tuning the large model.

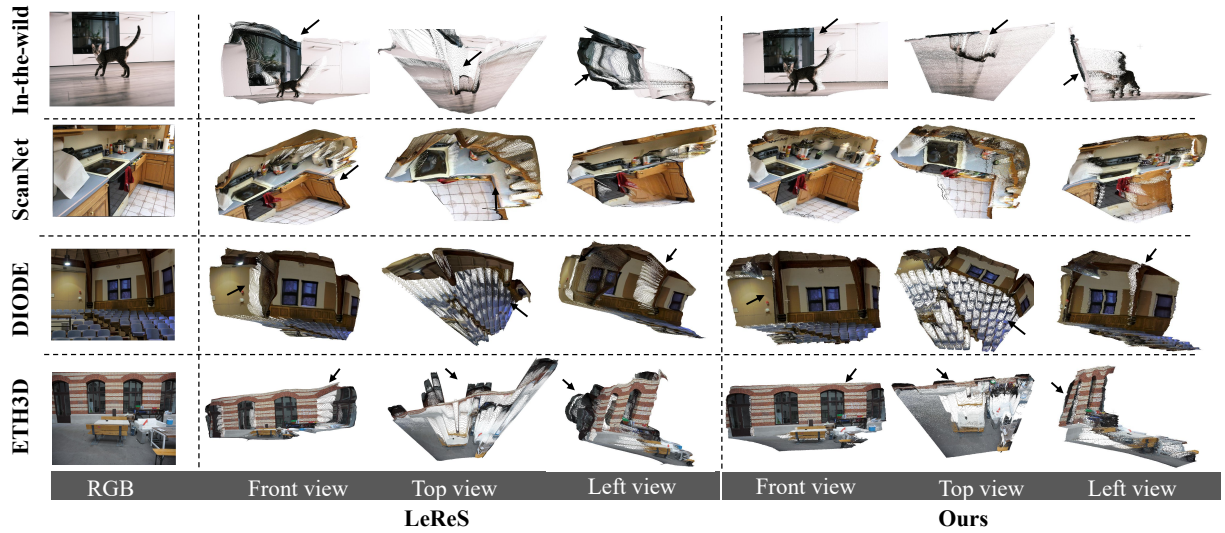

Figure 5: Qualitative comparison of 3D reconstruction. We compare with LeReS (Yin et al. 2021) across diverse scenes. The black arrow highlights the area of inconsistent geometry that has been reconstructed by alternative methodologies.

Methods	Waymo		Scenes11	
	e_f	e_b	e_f	e_b
DiffCalib (w/o pre-train)	0.263	0.062	0.214	0.034
DiffCalib	**0.188**	**0.054**	**0.108**	**0.029**

Table 4: Ablation study for the pre-trained parameters of the U-Net. We observe that the U-Net pre-training parameters can enhance the performance and facilitate the convergence.

Methods	Ensemble Times	Waymo		RGBD		Scenes11	
		e_f	e_b	e_f	e_b	e_f	e_b
w/o ensemble	0	0.160	0.098	0.141	0.068	0.201	0.079
ensemble	10	**0.145**	**0.053**	**0.084**	**0.040**	**0.176**	**0.038**

Table 5: Ablation study of the ensemble process. With the ensemble process, our DiffCalib demonstrates significantly improved robustness in camera intrinsic estimation.

Ensemble of Incidence Noise. Using the diffusion model to generate the incident map offers significant variety. By denoising multiple noise samples and averaging the results, the confidence in the predicted incident map increases. The results in Table 5 demonstrates this benefit, showing that ensemble noise improves results across several in-the-wild datasets. This aligns with real-world scenarios where images with slight variations may share the same incident map.

The Utilization of Incident Maps Improves Depth Estimation. For depth estimation, we evaluated our model on the NYU (Silberman et al. 2012) and ETH3D (Schops et al. 2017) datasets using affine-invariant settings. Trained only on indoor scenes, our model was tested on NYU's indoor scenes and ETH3D's mixed indoor and outdoor scenes. Table 6 shows that incorporating both depth and incident maps enhances intrinsic estimation and depth performance. A model trained with both maps outperforms one trained with depth alone. Both models, trained on the Hypersim dataset, differ only in incident map supervision, and our evaluation aligns with other depth estimation methods to validate the results.

Methods	NYU(indoor)		ETH3D	
	$AbsRel \downarrow$	$\delta 1 \uparrow$	$AbsRel \downarrow$	$\delta 1 \uparrow$
DiffCalib (w/o incident)	0.086	0.924	0.101	0.903
DiffCalib	**0.082**	**0.927**	**0.086**	**0.918**

Table 6: Ablation for the influence of incident map estimation to depth estimation. DiffCalib (w/o incident) refers to training using only depth maps.

Qualitative Comparison with 3D Reconstruction. We conduct a comparative analysis in Figure 5 between our single image 3D reconstruction method and the recent LeReS approach (Yin et al. 2021) on ScanNet, DIODE, ETH3D dataset, and in-the-wild condition, both of which rely solely on a single image for 3D reconstruction. our method consistently outperforms LeReS in terms of both detail preservation and geometric accuracy.

Conclusion

In this paper, we present a diffusion-based approach, namely DiffCalib, for monocular camera calibration with a single in-the-wild image. We reformulate the calibration problem as a diffusion-based dense incident map generation task. By leveraging the successful visual prior knowledge transferring from image generation to incidence and depth estimation, our approach enables more robust and accurate in-the-wild camera calibration, showing superior generalization ability and effectiveness upon existing methods. Beyond calibration, we show that the robust performance of our DiffCalib can greatly benefit downstream applications such as 3D scene reconstruction from a single image.

Acknowledgements

This work is partially supported by the National Key R&D Program of China(No. 2022ZD0160101), the Natural Science Foundation of Zhejiang Province (No. LZ25F020008, No. LMS25F020004), and the National Natural Science Foundation of China(No. 62102364, No. 6240070772, No. 62206244).

References

Ahmadyan, A.; Zhang, L.; Ablavatski, A.; Wei, J.; and Grundmann, M. 2021. Objectron: A large scale dataset of object-centric videos in the wild with pose annotations. In *CVPR*.

Akinlar, C.; and Topal, C. 2011. EDLines: A real-time line segment detector with a false detection control. *Pattern Recognition Letters*.

Bae, G.; Budvytis, I.; and Cipolla, R. 2021. Estimating and exploiting the aleatoric uncertainty in surface normal estimation. In *ICCV*.

Baruch, G.; Chen, Z.; Dehghan, A.; Dimry, T.; Feigin, Y.; Fu, P.; Gebauer, T.; Joffe, B.; Kurz, D.; Schwartz, A.; and Shulman, E. 2021. ARKitScenes - A Diverse Real-World Dataset for 3D Indoor Scene Understanding Using Mobile RGB-D Data. In *NeurIPS Datasets and Benchmarks Track*.

Caesar, H.; Bankiti, V.; Lang, A.; Vora, S.; Liong, V.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; and Beijbom, O. 2020. nuScenes: A multimodal dataset for autonomous driving. In *CVPR*.

Chang, A.; Funkhouser, T.; Guibas, L.; Hanrahan, P.; Huang, Q.; Li, Z.; Savarese, S.; Savva, M.; Song, S.; Su, H.; Xiao, J.; Yi, L.; and Yu, F. 2015. Shapenet: An information-rich 3D model repository. *arXiv preprint arXiv:1512.03012*.

Chen, B.; Chin, T.-J.; and Li, N. 2019. BPnP: Further empowering end-to-end learning with back-propagatable geometric optimization. *arXiv: 1909.06043*.

Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; and Schiele, B. 2016. The cityscapes dataset for semantic urban scene understanding. In *CVPR*.

Coughlan, J. M.; and Yuille, A. L. 1999. Manhattan world: Compass direction from a single image by bayesian inference. In *ICCV*.

Dai, A.; Chang, A. X.; Savva, M.; Halber, M.; Funkhouser, T.; and Nießner, M. 2017. ScanNet: Richly-annotated 3D Reconstructions of Indoor Scenes. In *CVPR*.

Fischler, M. A.; and Bolles, R. C. 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6): 381–395.

Fu, X.; Yin, W.; Hu, M.; Wang, K.; Ma, Y.; Tan, P.; Shen, S.; Lin, D.; and Long, X. 2024. GeoWizard: Unleashing the Diffusion Priors for 3D Geometry Estimation from a Single Image. *arXiv preprint arXiv:2403.12013*.

Fuhrmann, S.; Langguth, F.; and Goesele, M. 2014. Mve-a multi-view reconstruction environment. In *GCH*.

Geiger, A.; Lenz, P.; Stiller, C.; and Urtasun, R. 2013. Vision meets robotics: The KITTI dataset. *IJRR*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A. C.; and Bengio, Y. 2014. Generative Adversarial Networks. *Communications of the ACM*.

Grabner, A.; Roth, P.; and Lepetit, V. 2019. GP^{2}C: Geometric projection parameter consensus for joint 3D pose and focal length estimation in the wild. In *ICCV*.

He, W.; Fu, S.; Liu, M.; Wang, X.; Xiao, W.; Shu, F.; Wang, Y.; Zhang, L.; Yu, Z.; Li, H.; et al. 2025. Mars: Mixture of auto-regressive models for fine-grained text-to-image synthesis. In *AAAI*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. arXiv:2006.11239.

Hold-Geoffroy, Y.; Sunkavalli, K.; Eisenmann, J.; Fisher, M.; Gambaretto, E.; Hadap, S.; and Lalonde, J.-F. 2018. A perceptual measure for deep single image camera calibration. In *CVPR*.

Hu, M.; Brazil, G.; Li, N.; Ren, L.; and Liu, X. 2023. Camera Self-Calibration Using Human Faces. In *FG*.

Jin, L.; Zhang, J.; Hold-Geoffroy, Y.; Wang, O.; Blackburn-Matzen, K.; Sticha, M.; and Fouhey, D. F. 2023. Perspective Fields for Single Image Camera Calibration. In *CVPR*.

Jin, L.; Zhang, J.; Hold-Geoffroy, Y.; Wang, O.; Matzen, K.; Sticha, M.; and Fouhey, D. 2022. Perspective Fields for Single Image Camera Calibration.

Kasiopy. 2023. Multi-Resolution Noise for Diffusion Model Training. https://wandb.ai/johnowhitaker/multires_noise/reports/Multi-Resolution-Noise-for-Diffusion-Model-Training--VmlldzozNjYyOTU2?s=31. Last accessed 17.11.2023.

Ke, B.; Obukhov, A.; Huang, S.; Metzger, N.; Daudt, R. C.; and Schindler, K. 2023. Repurposing Diffusion-Based Image Generators for Monocular Depth Estimation. *arXiv preprint arXiv: 2312.02145*.

Lee, J.; Go, H.; Lee, H.; Cho, S.; Sung, M.; and Kim, J. 2021. Ctrl-C: Camera calibration transformer with line-classification. In *ICCV*.

Lee, J.; Sung, M.; Lee, H.; and Kim, J. 2020. Neural geometric parser for single image camera calibration. In *ECCV*.

Liu, M.; Ma, Y.; Zhen, Y.; Dan, J.; Yu, Y.; Zhao, Z.; Hu, Z.; Liu, B.; and Fan, C. 2025a. Llm4gen: Leveraging semantic representation of llms for text-to-image generation. In *AAAI*.

Liu, M.; Wu, F.; Li, B.; Lu, Z.; Yu, Y.; and Li, X. 2025b. Envisioning Class Entity Reasoning by Large Language Models for Few-shot Learning. In *AAAI*.

Ranftl, R.; Bochkovskiy, A.; and Koltun, V. 2021. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, 12179–12188.

Ranftl, R.; Lasinger, K.; Hafner, D.; Schindler, K.; and Koltun, V. 2022. Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-Shot Cross-Dataset Transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3).

Roberts, M.; Ramapuram, J.; Ranjan, A.; Kumar, A.; Bautista, M. A.; Paczan, N.; Webb, R.; and Susskind, J. M. 2021. Hypersim: A Photorealistic Synthetic Dataset for Holistic Indoor Scene Understanding. In *ICCV*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*, 10684–10695.

Schops, T.; Schonberger, J. L.; Galliani, S.; Sattler, T.; Schindler, K.; Pollefeys, M.; and Geiger, A. 2017. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *CVPR*, 3260–3269.

Silberman, N.; Hoiem, D.; Kohli, P.; and Fergus, R. 2012. Indoor segmentation and support inference from RGBD images. In *ECCV*, 746–760.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Sturm, J.; Engelhard, N.; Endres, F.; Burgard, W.; and Cremers, D. 2012. A benchmark for the evaluation of RGBD SLAM systems. In *IROS*.

Sturm, P. 2005. Multi-view geometry for general camera models. In *CVPR*.

Sun, P.; Kretzschmar, H.; Dotiwalla, X.; Chouard, A.; Patnaik, V.; Tsui, P.; Guo, J.; Zhou, Y.; Chai, Y.; Caine, B.; Vasudevan, V.; Han, W.; Ngiam, J.; Zhao, H.; Timofeev, A.; Ettinger, S.; Krivokon, M.; Gao, A.; Joshi, A.; Zhang, Y.; Shlens, J.; Chen, Z.; and Anguelov, D. 2020. Scalability in perception for autonomous driving: Waymo open dataset. In *CVPR*.

Von Gioi, R.; Jakubowicz, J.; Morel, J.-M.; and Randall, G. 2008. LSD: A fast line segment detector with a false detection control. *TPAMI*.

Xiao, J.; Owens, A.; and Torralba, A. 2013. Sun3D: A database of big spaces reconstructed using SFM and object labels. In *ICCV*.

Xu, G.; Ge, Y.; Liu, M.; Fan, C.; Xie, K.; Zhao, Z.; Chen, H.; and Shen, C. 2024. Diffusion Models Trained with Large Data Are Transferable Visual Models. *arXiv preprint arXiv:2403.06090*.

Xu, G.; Yin, W.; Chen, H.; Shen, C.; Cheng, K.; Wu, F.; and Zhao, F. 2022. Towards 3d scene reconstruction from locally scale-aligned monocular video depth. *arXiv preprint arXiv:2202.01470*.

Yin, W.; Zhang, C.; Chen, H.; Cai, Z.; Yu, G.; Wang, K.; Chen, X.; and Shen, C. 2023. Metric3D: Towards zero-shot metric 3d prediction from a single image. In *ICCV*.

Yin, W.; Zhang, J.; Wang, O.; Niklaus, S.; Mai, L.; Chen, S.; and Shen, C. 2021. Learning to Recover 3D Scene Shape from a Single Image. In *CVPR*.

Yu, X.; Xu, M.; Zhang, Y.; Liu, H.; Ye, C.; Wu, Y.; Yan, Z.; Zhu, C.; Xiong, Z.; Liang, T.; Chen, G.; Cui, S.; and Han, X. 2023. MVImgNet: A Large-scale Dataset of Multi-view Images. In *CVPR*.

Yuan, W.; Gu, X.; Dai, Z.; Zhu, S.; and Tan, P. 2022. NeWCRFs: Neural Window Fully-connected CRFs for Monocular Depth Estimation. In *CVPR*.

Zhang, D.; Zhang, H.; Tang, J.; Hua, X.; and Sun, Q. 2020a. Causal Intervention for Weakly-Supervised Semantic Segmentation. *Neural Information Processing Systems*.

Zhang, D.; Zhang, H.; Tang, J.; Wang, M.; Hua, X.; and Sun, Q. 2020b. Feature Pyramid Transformer. *European Conference on Computer Vision*.

Zhang, Z. 2000. A flexible new technique for camera calibration. *TPAMI*.

Zhu, S.; Kumar, A.; Hu, M.; and Liu, X. 2023. Tame a Wild Camera: In-the-Wild Monocular Camera Calibration. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.