BOOST 3D RECONSTRUCTION USING DIFFUSION-BASED MONOCULAR CAMERA CALIBRATION

Anonymous authors
Paper under double-blind review

ABSTRACT

In this paper, we present *DM-Calib*, a diffusion-based approach for estimating pinhole camera intrinsic parameters from a single input image. Monocular camera calibration is essential for many 3D vision tasks. However, most existing methods depend on handcrafted assumptions or are constrained by limited training data, resulting in poor generalization across diverse real-world images. Recent advancements in stable diffusion models, trained on massive data, have shown the ability to generate high-quality images with varied characteristics. Emerging evidence indicates that these models implicitly capture the relationship between camera focal length and image content. Building on this insight, we explore how to leverage the powerful priors of diffusion models for monocular pinhole camera calibration. Specifically, we introduce a new image-based representation, termed Camera Image, which losslessly encodes the numerical camera intrinsics and integrates seamlessly with the diffusion framework. Using this representation, we reformulate the problem of estimating camera intrinsics as the generation of a dense Camera Image conditioned on an input image. By fine-tuning a stable diffusion model to generate a Camera Image from a single RGB input, we can extract camera intrinsics via a RANSAC operation. We further demonstrate that our monocular calibration method enhances performance across various 3D tasks, including zero-shot metric depth estimation, 3D metrology, pose estimation and sparse-view reconstruction. Extensive experiments on multiple public datasets show that our approach significantly outperforms baselines and provides broad benefits to 3D vision tasks.

1 INTRODUCTION

Camera calibration is a foundational task in 3D computer vision, critical for numerous applications such as camera pose estimation (Schönberger & Frahm, 2016), 3D reconstruction (Seitz et al., 2006), and zero-shot metric depth estimation (Yin et al., 2023). Traditional methods primarily focus on multi-view calibration, typically involving multiple images of fixed intrinsics (Pollefeys & Gool, 1997) or multiple images of checkerboard patterns (Zhang, 2000). However, these methods depend heavily on dense multi-view images with sufficient overlap, making them cumbersome and often impractical for sparse-view or even monocular setups. Consequently, monocular camera calibration has garnered significant research interest.

Monocular camera calibration is inherently an ill-posed problem, requiring additional information to address it. Traditional approaches have attempted to incorporate handcrafted knowledge, such as the gravity direction (Veicht et al., 2024), Manhattan World constraints (Liu & Cui, 2023), and human face priors (Hu et al., 2023). However, these handcrafted insights often fail to generalize effectively across diverse real-world scenarios. To overcome these limitations, recent studies (Zhu et al., 2023) recast monocular camera calibration as a learning-based regression problem, leveraging a single image to directly infer its intrinsic parameters.

While learning-based methods benefit from data-driven knowledge, outperforming traditional approaches, they are constrained by the limited availability of public datasets. As a result, these methods tend to overfit on training data and exhibit poor generalization to unseen scenarios. This limitation raises a critical question: *What kind of knowledge is necessary to develop a robust camera calibration method that exhibits strong generalization capabilities?*

One promising solution lies in leveraging stable diffusion priors (Rombach et al., 2022). The intuition behind stems from a key observation: stable diffusion models possess an implicit understanding of imaging across different focal lengths. As is widely known, *Cameras with long focal lengths tend to compress spatial relationships, resulting in a more flattened image perspective, while wide-angle cameras exaggerate depth and distance, resulting in a more pronounced perspective effect.* As illustrated in Fig. 1, we present two portrait images generated by a stable diffusion model (Rombach et al., 2022) using similar text prompts but with varying focal length descriptions. Clearly, the left image prompted with "long focal length", exhibits a more blurred background and shallower depth of field compared to the right image. This demonstrates that, by training on large-scale image-text pairs, these models encapsulate knowledge related to imaging characteristics associated with different focal lengths.

(a). Long focal length (b). Short focal length

Figure 1: Images generated using text prompts that specify different focal lengths.

Despite these advancements, a key challenge persists: how to effectively leverage diffusion priors for high-precision camera calibration? In this paper, we set out to explore this question and introduce ***DM-Calib***, a diffusion-based model for estimating intrinsic camera parameters from a single image. We recognize that the representation format used to encode camera intrinsics is essential for effective monocular camera calibration using diffusion models. To address this, we conduct in-depth investigation for various camera representations and develop the Camera Image, a novel image-based representation specifically engineered for seamlessly integration with pre-trained diffusion models, thereby preventing loss of information. Subsequently, we train a diffusion model that takes a single image as input and generates the Camera Image, followed by a RANSAC algorithm to solve the camera intrinsic parameters. Moreover, we demonstrate how to integrate the proposed camera calibration with diffusion-based metric depth estimation, which allows the recovery of true-scale depth measurements from a single image. Furthermore, our experiments show that the recovered camera calibration results significantly improve the performance of various downstream tasks, including camera pose estimation, sparse-view 3D reconstruction, and novel view synthesis, showcasing the robustness and effectiveness of *DM-Calib* in advancing accurate monouclar camera calibration.

To summarize, our main contributions are:

- We introduce the Camera Image, a novel image-based representation specifically designed to encode camera intrinsic information, optimized to use with pretrained diffusion models.
- We present ***DM-Calib***, a generative foundation model that provides highly accurate estimations of camera intrinsics. Additionally, it can seamlessly integrate with various downstream tasks, showcasing its effectiveness and robustness to images from various scenarios.
- Extensive experiments on multiple public datasets show that our approach significantly outperforms baselines and provides broad benefits to 3D vision tasks.

2 Related Work

2.1 Monocular Camera Calibration

Calibrating camera intrinsics, also known as self-calibration, is one of the most fundamental problems in geometric computer vision. Geometric methods typically assume multiple input images with the same intrinsic matrix. These methods can be broadly classified as direct methods and stratified methods. Direct methods (Zeller & Faugeras, 1996; Hartley, 1997; Luong & Faugeras, 1997) solve the Kruppa's equation (Gallego et al., 2018) to obtain camera intrinsic parameters, which are generally more fragile to noises. In comparison, stratified methods (Hartley, 1993; Triggs, 1997; Pollefeys & Gool, 1997) estimate camera intrinsics from a projective reconstruction by gradually recovering the affine and Euclidean structures such as the plane at infinity and absolute quadrics.

Traditional monocular camera calibration methods typically assume specific geometric structures to estimate intrinsics. For example, with Manhattan World assumption (Coughlan & Yuille, 1999), camera intrinsics can be inferred from vanishing points (Lee et al., 2013; Schindler et al., 2004;

Wildenauer & Hanbury, 2012), or by jointly estimating the horizon line (Zhai et al., 2016; Simon et al., 2018). Other approaches rely on calibration objects such as checkerboards (Zhang, 2000), line segments (Zhang et al., 2016), spheres (Zhang et al., 2007), pyramid frustums (Jiang et al., 2009), or even human faces (Hu et al., 2023; Liu & Cui, 2023). Despite producing satisfactory results, these methods are constrained by their reliance on specific objects, which limits their applicability in unconstrained, real-world scenarios. Recently, Zhu et al. (2023) introduced incident maps to regress camera intrinsics, enabling the detection of geometric manipulations like cropping.

Apart from the above geometry-based methods, some works attempt to leverage the strong generative models for camera calibration. To the best of our knowledge, the approach by He et al. (2024) is the only existing method that formulates camera intrinsic estimation as a generative task, leveraging incident maps with diffusion models. Although this method outperforms non-diffusion-based methods such as Zhu et al. (2023) in unconstrained settings, we argue that the full potential of diffusion models remains under-utilized and that its performance hinges on joint training with depth images. In contrast, our approach introduces a camera representation that is more inherently compatible with diffusion models, thereby eliminating the need to generate non-textured incident maps or rely on the joint training of additional geometric information.

2.2 Diffusion Models in 3D tasks

Recently, diffusion models (Ho et al., 2020; Song et al., 2020) have emerged as powerful tools across various domains, particularly in the field of computer vision, where Text-to-Image diffusion models (Saharia et al., 2022b) and their extensions (Poole et al., 2022; Rombach et al., 2022; Saharia et al., 2022a; Zhang et al., 2023) have garnered significant attention. Compared to GAN-based approaches (Bhattad et al., 2024), several studies have highlighted the advantages of diffusion models, especially when used as prior geometric cues in 3D tasks. Notable examples include view synthesis (Liu et al., 2023a; Long et al., 2024), camera calibration (He et al., 2024), normal estimation (Ye et al., 2024; Liu et al., 2023b), and depth estimation (Fu et al., 2024; Ke et al., 2024). In this work, we focus on leveraging diffusion models for camera intrinsic estimation, which serves as a foundation for enhancing a series of downstream tasks, such as monocular metric depth estimation, pose estimation, and sparse-view reconstruction.

2.3 Monocular Depth Estimation

For depth estimation, several works (Ke et al., 2024; Fu et al., 2024; Hu et al., 2024b) have shown that diffusion models can be fine-tuned to predict affine-invariant depth, achieving not only finer detail but also more accurate estimates than traditional methods (Ranftl et al., 2020; Yin et al., 2021; Yang et al., 2024; Yin et al., 2022). Recent studies (Xu et al., 2024; Ye et al., 2024; Garcia et al., 2024) have highlighted that diffusion models can serve as pre-trained networks for deterministic, one-step inference. However, none of the current approaches leverage pre-trained diffusion models for metric depth prediction. Existing zero-shot monocular metric depth estimation methods, such as (Bhat et al., 2023; Yin et al., 2023; Piccinelli et al., 2024; Hu et al., 2024a), have demonstrated accurate results, yet they still face challenges in capturing geometric details and foreground-background relationships, particularly in outdoor environments. Moreover, these methods usually employ contrastive learning pretrained encoders (i.e., DINO (Caron et al., 2021)) or classification pretrained encoders (Deng et al., 2009), and randomly initialize the decoder, which are trained on fewer images compared to diffusion models. Building upon this, we extend diffusion models to metric depth estimation. To the best of our knowledge, our approach is the first to employ pre-trained diffusion models for metric depth estimation, achieving finer geometric detail and competitive performance across diverse benchmarks.

3 Method

Given a single input image $\mathbf{x} \in \mathbb{R}^{H \times W \times 3}$, our objective is to recover its camera intrinsic matrix $\boldsymbol{K}$. To efficiently and losslessly integrate camera intrinsics prediction with diffusion models (Rombach et al., 2022), we introduce Camera Image (Fig. 3) to encode camera intrinsics as a detail-preserving color image (see Sec. 3.2). We reformulate camera calibration as a conditional generation task, transforming the text-to-image (T2I) diffusion model into an image-to-Camera-Image (I2C) model, from which camera intrinsics are recovered via a RANSAC algorithm (see Sec. 3.3). Our proposed calibration method significantly boosts the performance of downstream 3D vision tasks, including

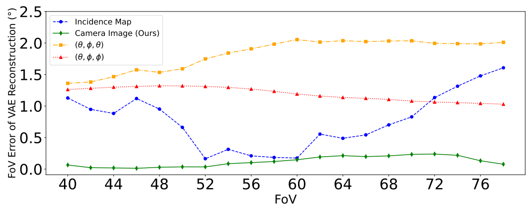
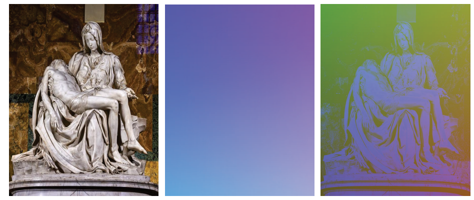


Figure 2: **Error analysis of camera representations.** We first use pre-trained VAE to encode and decode each camera representation, and plot the FoV reconstruction errors (°) here.



a). RGB b). Incidence Map c). Camera Image

Figure 3: **Visualization of incidence map and Camera Image.** We show the input RGB image, the incidence map and our proposed Camera Image for reference.

monocular metric depth estimation, 3D reconstruction, and pose estimation (see Sec. 3.4). Before introducing our method, we first revisit the preliminary concepts related to diffusion models.

3.1 Preliminaries on Diffusion Model

Diffusion models (DMs) (Ho et al., 2020) learn to model a data distribution $p_{\text{data}}(\mathbf{x})$ by progressively denoising a noise variable that is initially sampled from a normal distribution. Recognizing the efficiency issue associated with generating high-resolution images, Rombach et al. (2022) introduced latent diffusion models (LDMs), which operate the diffusion process in the latent space of a pretrained variational autoencoder (VAE) (Kingma, 2013) with an encoder $\mathcal{E}$ and a decoder $\mathcal{D}$.

For any given input image $\mathbf{x}$, the corresponding latent code is generated by the VAE encoder: $\mathbf{z} = \mathcal{E}(\mathbf{x})$. The forward diffusion process incrementally adds noise to these latents following $\mathbf{z}_t := \alpha_t \mathbf{z} + \sigma_t \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, and α_t and σ_t are parameters defined by the noise schedule, with $t \sim p_t$ representing the time step within the diffusion schedule. The denoising network, denoted as $\boldsymbol{\epsilon}_\theta$, aims to reverse the diffusion process to recover the noise-free latent code $\hat{\mathbf{z}}$ from the final noisy latent code $\mathbf{z}_T$. This is achieved by predicting the noise component $\boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t)$ at each diffusion step. The original image $\mathbf{x}$ is then reconstructed from this denoised latent code using the VAE decoder as $\hat{\mathbf{x}} = \mathcal{D}(\hat{\mathbf{z}})$. The whole diffusion model is optimized by minimizing the denoising score matching objective, defined as follows: $\mathbb{E}_{\mathbf{z},\boldsymbol{\epsilon},t}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t)\|_2^2\right]$. This objective measures the squared Euclidean distance between the actual noise $\boldsymbol{\epsilon}$ and the predicted noise. By minimizing this objective, the denoising network learns to accurately estimate the noise, thereby effectively reversing the diffusion process and reconstructing the original data distribution.

3.2 Camera Image Representation

Monocular camera calibration aims to recover the camera intrinsics matrix $\boldsymbol{K}$, which typically composed of four parameters: f_x, f_y, c_x, and c_y, corresponding to the focal lengths and the optical center coordinates along the x-axis and y-axis, respectively. This numeric-based representation, however, does not align well with image-based diffusion models, which are primarily designed for generating spatial images. The challenge, therefore, becomes how to effectively leverage powerful pretrained SD models to retrieve implicit camera information.

To address this challenge, we propose a novel image-based representation, called "Camera Image", which encodes the camera intrinsic parameters into a 3-channel color image (refer to Fig. 3 for visual representation). This representation seamless integrates with existing diffusion models with minimal architecture modifications. We reformulate the camera intrinsics into a 2-channel pseudo-spherical representation defined by azimuth $\boldsymbol{\theta}$ and elevation $\boldsymbol{\varphi}$. This two-channel formulation enables us to explore the choice of the third channel to prevent mode collapse. Given that VAE encoders typically take a three-channel image as input, it is crucial to determine how to effectively fill the third channel. Simply duplicating one of the existing channels $\boldsymbol{\theta}, \boldsymbol{\varphi}$ or adding a constant value channel leads to suboptimal results, as detailed in Fig. 2. To enhance the camera representation, we propose a simple yet effective solution by incorporating the grayscale image $\mathbf{g}$ of the input $\mathbf{x}$ into the dense camera representation, reducing the domain gap between the input images and those generated by diffusion models. Consequently, our proposed camera image $\mathbf{c}$ is defined as follows,

$$\mathbf{c}_{(u,v)} = \left[\arctan\left(\frac{r_1}{r_3}\right), \arccos\left(r_2\right), \mathbf{g}_{(u,v)}\right], \tag{1}$$

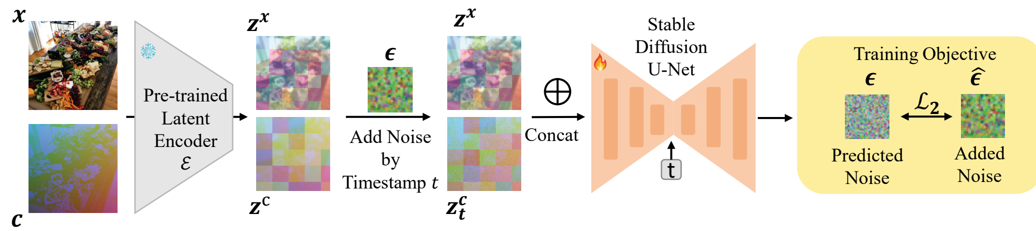


Figure 4: **The overview training framework of *DM-Calib*.** The input image $\mathbf{x}$ and the camera image $\mathbf{c}$ are first encoded into latent space using a frozen VAE encoder. We then inject timestamp-relevant noise $\boldsymbol{\epsilon}$ into the camera's latent code, which is concatenated with the image latent code and fed into the subsequent UNet. The UNet is fine-tuned to predict the added noise $\hat{\boldsymbol{\epsilon}}$.

where $\vec{\boldsymbol{r}} = [r_1, r_2, r_3] \cong \boldsymbol{K}^{-1}[u, v, 1]^T$, $\boldsymbol{K}$ is the intrinsic matrix, and (u, v) are the pixel coordinates. $\vec{\boldsymbol{r}}$ is normalized as a unit vector, and $\mathbf{g}_{(u,v)}$ is the gray-scale pixel value sampled at coordinate (u, v). As shown on the right side of Fig. 3, the proposed camera image preserves the high-frequency details of the original scene, making it closely resemble real-world images that diffusion models are designed to process. The incidence map (shown in the middle of Fig. 3), proposed by concurrent research (He et al., 2024), however, exhibits a large domain gap from the original image domain, which results in suboptimal intrinsics estimations according to our experimental results in Sec. 4.2.

3.3 Camera Intrinsic Estimation

Diffusion Model for Camera Image Prediction. Our camera intrinsic estimation is built upon the pre-trained latent diffusion model, Stable Diffusion v2.1 (Rombach et al., 2022), which leverages robust image priors trained on the billion-scale LAION-5B dataset (Schuhmann et al., 2022). To facilitate the generation of the proposed camera image, we detail our training pipeline in Fig. 4. First, a frozen VAE $\mathcal{E}$ encodes the RGB image $\mathbf{x}$ and its corresponding camera image $\mathbf{c}$ into latent space $\mathbf{z}^{\boldsymbol{x}}$ and $\mathbf{z}^{\mathbf{c}}$, respectively. Multi-resolution noise (Kasiopy, 2023) $\boldsymbol{\epsilon}^{\mathbf{c}}$ is then added to the camera latents $\mathbf{z}^{\mathbf{c}}$, forming the noisy code $\mathbf{z}^{\mathbf{c}}_T$. This code is concatenated with $\mathbf{z}^{\boldsymbol{x}}$, serving as the input for the pretrained U-Net. To accommodate our inputs, we double the input channels of the original U-Net and adjust the corresponding parameter weights accordingly. The U-Net is targeted to predict the added noise, and the final loss function is expressed as:

$$\mathcal{L} = \mathbb{E}_{\mathbf{x},\mathbf{c}\sim p_{\text{data}},t\sim p_t,\boldsymbol{\epsilon}^{\mathbf{c}}} \left\| \hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}}(\mathbf{z}^{\mathbf{c}}_t; \mathbf{z}^{\boldsymbol{x}}) - \mathbf{v}_t \right\|_2^2 , \tag{2}$$

where $\mathbf{v}_t = \alpha_t \boldsymbol{\epsilon}^{\mathbf{c}}_t - \beta_t \mathbf{z}^{\mathbf{c}}_t$, and $\boldsymbol{\epsilon}^{\mathbf{c}}_t$ is the sampled multiscale noise for the camera image. During inference, we can formulate the generation of the camera image within a generative framework $f\colon \boldsymbol{x} \in \mathbb{R}^3 \to \hat{\boldsymbol{c}} \in \mathbb{R}^3$ utilizing v-prediction (Salimans & Ho, 2022), as follows:

$$f(\mathbf{z}^{\boldsymbol{x}}) = p(\hat{\mathbf{z}}^{\mathbf{c}}_T) \prod_{t=1}^{T} p_{\boldsymbol{\theta}} \left(\hat{\mathbf{z}}^{\mathbf{c}}_{t-1} | \hat{\mathbf{z}}^{\mathbf{c}}_t \right) , \tag{3}$$

where $\mathbf{z}$ is the latent feature and $\hat{\mathbf{z}}^{\mathbf{c}}_T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$. After completing the multi-step denoising process using the U-Net, the denoised camera latent representation $\hat{\mathbf{z}}^{\mathbf{c}}$ is sent to the frozen VAE decoder, yielding the final camera image $\hat{\mathbf{c}} = \mathcal{D}(\hat{\mathbf{z}}^{\mathbf{c}})$, and we verified that our camera image provides negligible error with respect to the VAE encoder-decoder reconstruction (see Fig. 2). From this generated image, we can extract the numerical representation of the camera intrinsic parameters.

Recover Camera Intrinsics From Camera Image. With the recoverd camera image $\hat{\mathbf{c}}$, camera intrinsic matrix $\boldsymbol{K}$ can then be inferred from the first two channels of the camera image via the relation between camera image and camera intrinsic $\boldsymbol{K}$ in Eq. 1 :

$$\tan(c_\theta) f_x + c_x = u, \qquad \frac{1}{\cos(c_\theta)\tan(c_\varphi)} f_y + c_y = v, \tag{4}$$

where $\hat{\mathbf{c}}_{(u,v)} = [c_\theta, c_\varphi, g]$ presents the pixel value of the camera image. Since, every two pixels can be used to solve the camera intrinsics, we employ the RANSAC algorithm to align the two lines using all pixel in the camera image. Here, the focal length $f_{x/y}$ and the optical center $c_{x/y}$ are represented as the slope and intercept of the best-fit line of Eq. 4, respectively.

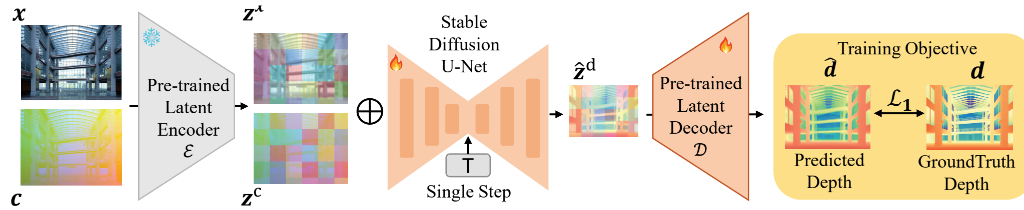


Figure 5: **The overview of metric depth training pipeline.** The encoded image and camera image z^x and z_c are concatenated and sent to pretrained U-Net. Then we employ single-step diffusion at timestamp T to generate depth latent code $\hat{z}_d$, which is then decoded into predicted metric depth $\hat{d}$.

Table 1: **Monocular Camera Calibration on Zero-Shot Datasets.** We report the calibration errors for both focal length and optical center. †: focuses on focal length prediction. ‡: Waymo and ScanNet are in the training set. ⋇: joint training with depth.

Method	Waymo		RGBD		ScanNet		MVS		Scenes11		Average	
	e_f	e_b	e_f	e_b	e_f	e_b	e_f	e_b	e_f	e_b	e_f	e_b
Perspective †	0.444	-	0.166	-	0.189	-	0.185	-	0.211	-	0.239	-
GeoCalib †	0.285	-	0.203	-	0.137	-	0.104	-	0.344	-	0.215	-
WildCame	0.210	0.053	0.097	0.039	0.128	0.041	0.170	0.028	0.170	0.044	0.155	0.041
DiffCalib	0.188	0.053	0.092	0.018	0.089	0.041	0.135	0.032	0.108	0.029	0.122	0.030
DiffCalib-D ⋇	0.145	0.053	0.084	0.040	**0.055**	0.036	0.108	0.036	0.176	0.038	0.095	0.041
Unidepth ‡	-	-	0.055	0.052	-	-	0.482	**0.001**	0.510	0.051	0.350	0.030
Ours	**0.115**	**0.036**	**0.041**	**0.010**	0.089	**0.024**	**0.087**	0.008	**0.061**	**0.010**	**0.078**	**0.017**

With the proposed camera image and intrinsics estimation, our approach offers applicability to various 3D downstream tasks, including monocular metric depth estimation (MMDE), camera pose estimation, and 3D reconstruction.

3.4 DOWNSTREAM 3D VISION TASKS

Monocular Metric Depth Estimation. To predict metric depth from a single image, the model must possess a deep understanding of the image perspective and estimate accurate intrinsic parameters of the camera. By leveraging the proposed camera calibration method, we repurpose diffusion-based image generators for accurate metric depth estimation. Previous works (Ke et al., 2024; Fu et al., 2024) mainly investigate affine-invariant depth estimation. However, we find the VAE decoder $\mathcal{D}$ can only predict values in limited range, thus limiting the performance of metric depth estimation. To fix this issue, we formulate stochastic multi-step denoise SD model as one-step deterministic forward process as shown in Fig. 5. Specifically, we first encode RGB image and our designed camera image $\hat{\mathbf{c}}$ via VAE encoder into latent space, noting that no noise is added to both of the latent features. Then, the latent features are sent to the UNet to predict the latent depth features $\hat{\mathbf{z}}_{\boldsymbol{d}}$, and the final depth predictions $\hat{\boldsymbol{d}}$ are obtained via the decoder of the VAE. Note that both U-Net $\mathcal{U}$ and the VAE decoder $\mathcal{D}$ are trained to allow predictions in any range. Given the depth labels $\boldsymbol{d}$ with its sparse mask $\boldsymbol{M}$, the training loss is given by:

$$\mathcal{L}_{\text{depth}} = \mathbb{E}_{\mathbf{x},\hat{\mathbf{c}}\sim p_{\text{data}}} \|\boldsymbol{M} \odot [\mathcal{D}(\mathcal{U}(\mathbf{z}^{\boldsymbol{x}}, \hat{\mathbf{z}}^{\mathbf{c}})) - \boldsymbol{d}]\| \tag{5}$$

Sparse-View 3D Reconstruction & Pose Estimation. Capturing sparse-view images with varying camera settings, particularly focal lengths, complicates object reconstruction using structure-from-motion (SfM) methods like COLMAP (Schönberger & Frahm, 2016) due to missing intrinsic parameters and low image overlap. While approaches like DUST3R (Wang et al., 2024) optimize both intrinsic and extrinsic parameters for reconstruction from sparse viewpoints, they struggle with significantly different intrinsic settings. To address this, we incorporate our estimated intrinsics as a geometry cue for the subsequent reconstruction in Wang et al. (2024), fixing the focal length during optimization to demonstrate the robustness of our approach. In Wang et al. (2024), the pointmap $X \in \mathbb{R}^{H\times W\times 3}$ is predicted, and the relative pose $P^* = [R^*|t^*]$ can be recovered via Procrustes alignment (Luo & Hancock, 1999). More details are provided in our appendix.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets. For camera intrinsic estimation, the training data is sourced from a variety of datasets, including NuScenes (Caesar et al., 2020), KITTI (Geiger et al., 2012), CityScapes (Cordts et al., 2016), NYUv2 (Nathan Silberman & Fergus, 2012), SUN3D (Xiao et al., 2013), ARKitScenes (Baruch et al., 2021), Objectron (Ahmadyan et al., 2021), MVImgNet (Yu et al., 2023), Hypersim (Roberts et al., 2021), Virtual KITTI (Cabon et al., 2020), Taskonomy (Zamir et al., 2018), and TartanAir (Wang et al., 2020). We adopt Waymo (Sun et al., 2020a), RGBD (Sturm et al., 2012), ScanNet (Dai et al., 2017), MVS (Fuhrmann et al., 2014), and Scenes11 (Chang et al., 2015) datasets for zero-shot testing.

For metric depth training, we use Taskonomy (Zamir et al., 2018), Hypersim (Roberts et al., 2021), TartanAir (Wang et al., 2020), Virtual KITTI (Cabon et al., 2020), Waymo (Sun et al., 2020b) and Argoverse2 (Wilson et al., 2021). Additionally, we incorporate 10k synthetic city samples collected by ourselves. The evaluation is performed on NuScenes (Caesar et al., 2020), ETH3D (Schöps et al., 2017), Diode (Vasiljevic et al., 2019), VOID (Wong et al., 2020), IBims-1 (Koch et al., 2020), NYUV2 (Nathan Silberman & Fergus, 2012). For more details, please refer to the appendix.

Evaluation Protocols. For camera intrinsic estimation, we follow the evaluation protocol of (Zhu et al., 2023; He et al., 2024) using the relative error:

$$e_f = \max\left(\frac{|f'_x - f_x|}{f_x}, \frac{|f'_y - f_y|}{f_y}\right), e_b = \max\left(2 \cdot \frac{|c'_x - c_x|}{w}, 2 \cdot \frac{|c'_y - c_y|}{h}\right) \tag{6}$$

For depth estimation, we use the Absolute mean relative error (A.Rel), the percentage of inlier pixels δ_i with threshold 1.25^i and scale-invariant error in log scale $\mathrm{SI}_{\log} = 100\sqrt{\mathrm{Var}\left(\varepsilon_{\log}\right)}$.

Baselines. For camera calibration, we compare our method with three non-diffusion based methods, Jin et al. (2023), Zhu et al. (2023) Piccinelli et al. (2024) and Veicht et al. (2024), one diffusion-based method (He et al., 2024). For metric depth estimation, we compare our method with 4 state-of-the-art methods. For additional reference, we also evaluate the generated depth using affine-invariant depth protocols with several affine-invariant depth depth estimation methods.

Implementation Details. Our models are built on the pre-trained Stable Diffusion V2.1 model (Rombach et al., 2022). To train camera intrinsic estimation model, we employ the AdamW optimizer with a learning rate of $3e^{-5}$ and train the model for 30,000 iterations with a total batch size of 196 on a cluster of 8 Nvidia A800 GPUs. For metric depth estimation, we use the same optimizer and learning rate with a total batch size of 96, and the training process takes approximately 5 days to converge. For all of our downstream 3D vision tasks, we did not use the ground truth camera image but instead relied on intrinsic parameters predicted by our diffusion model.

4.2 CAMERA INTRINSIC EVALUATION

We first present our monocular camera calibration results on five zero-shot datasets in Tab. 1. As shown, our method achieves the highest calibration accuracy. Compared to the concurrent work DiffCalib (He et al., 2024), our approach performs better due to the superior suitability of the proposed Camera Image for diffusion priors, allowing seamless integration with stable diffusion models. Among the methods, Unidepth (Piccinelli et al., 2024) shows the weakest performance, particularly in real-world, unconstrained scenarios such as the MVS dataset. This subpar result may stem from unbalanced training between the tasks of metric depth estimation and camera intrinsic estimation. Geometry-inspired methods (Jin et al., 2023; Veicht et al., 2024) face challenges in achieving strong performance on these datasets as they heavily rely on geometric information, such as vanishing points. Notably, our method also performs well on the highly challenging Scenes11 dataset (Chang et al., 2015), which features randomly shaped, moving objects, further demonstrating its robustness in extreme conditions.

4.3 DEPTH EVALUATION

Metric Depth Comparison. We evaluate the performance of our method on metric depth estimation task. The quantitative and qualitative results are presented in Tab. 2 and Fig. 6 respectively. Our work

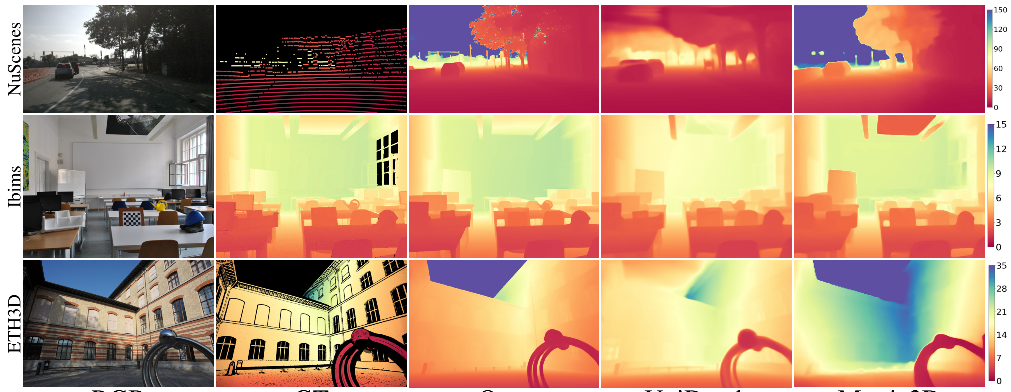


Figure 6: **Zero-Shot Metric Depth Estimation Results.** We present the predicted metric depth in both various scenes. Our method provides more detailed results and recovers accurate metric.

Table 2: **Comparison on Zero-Shot Metric Depth Evaluation.** We achieve comparable precision to state-of-the-art models while utilizing less training data.

Method	NYU-V2		NuScenes		ETH3D		DIODE (Indoor)		VOID		IBims-1	
	$\delta_1 \uparrow$	$\mathrm{SI}_{\log} \downarrow$	$\delta_1 \uparrow$	$\mathrm{SI}_{\log} \downarrow$	$\delta_1 \uparrow$	$\mathrm{SI}_{\log} \downarrow$	$\delta_1 \uparrow$	$\mathrm{SI}_{\log} \downarrow$	$\delta_1 \uparrow$	$\mathrm{SI}_{\log} \downarrow$	$\delta_1 \uparrow$	$\mathrm{SI}_{\log} \downarrow$
iDisc	93.8	8.85	39.4	37.1	35.6	27.5	23.8	15.8	55.3	20.3	48.9	13.2
ZoeDepth	90.1	–	28.3	31.5	35.0	17.6	36.9	12.8	63.4	15.9	58.0	10.9
Metric3D	92.6	9.13	72.3	29.0	45.6	18.9	39.2	11.1	65.9	16.2	79.7	10.1
UniDepth	**97.2**	**6.41**	83.3	22.9	22.9	13.1	**60.4**	**9.01**	**88.5**	**8.26**	79.4	8.88
Ours	93.1	8.35	**85.7**	**19.8**	**49.0**	**9.08**	42.2	13.3	73.1	15.3	**88.5**	**8.27**

Table 3: **Quantitative Comparison on 5 Zero-shot Affine-invariant Depth Benchmarks.** Despite targeting metric depth, we achieve performance comparable to SoTA affine-invariant depth methods.

Method	NYUv2		KITTI		ETH3D		ScanNet		DIODE-Full	
	AbsRel $\downarrow$	$\delta 1 \uparrow$	AbsRel $\downarrow$	$\delta 1 \uparrow$	AbsRel $\downarrow$	$\delta 1 \uparrow$	AbsRel $\downarrow$	$\delta 1 \uparrow$	AbsRel $\downarrow$	$\delta 1 \uparrow$
DiverseDepth (Yin et al., 2020)	11.7	87.5	19.0	70.4	22.8	69.4	10.9	88.2	37.6	63.1
MiDaS (Ranftl et al., 2022)	11.1	88.5	23.6	63.0	18.4	75.2	12.1	84.6	33.2	71.5
LeReS (Yin et al., 2021)	9.0	91.6	14.9	78.4	17.1	77.7	9.1	91.7	27.1	76.6
Omnidata v2 (Kar et al., 2022)	7.4	94.5	14.9	83.5	16.6	77.8	7.5	93.6	33.9	74.2
HDN (Zhang et al., 2022)	6.9	94.8	11.5	86.7	12.1	83.3	8.0	93.9	24.6	78.0
DPT (Ranftl et al., 2021)	9.8	90.3	10.0	90.1	7.8	94.6	8.2	93.4	**18.2**	75.8
Metric3D (Yin et al., 2023)	5.8	96.3	**5.8**	**97.0**	6.6	96.0	7.4	94.1	22.4	78.5
DepthAnything (Yang et al., 2024)	**4.3**	**98.1**	7.6	94.7	12.7	88.2	**4.2**	**98.0**	27.7	75.9
Marigold (Ke et al., 2024)	5.5	96.4	9.9	91.6	6.5	96.0	6.4	95.1	30.8	77.3
GeoWizard (Fu et al., 2024)	5.2	96.6	9.7	92.1	**6.4**	**96.1**	6.1	95.3	29.7	79.2
Ours	4.8	97.1	8.5	93.5	7.1	95.3	5.7	96.5	25.6	**79.4**

significantly outperforms strong baselines such as Metric3D (Yin et al., 2023) by a large margin, and achieves comparable performance with a concurrent work Unidepth (Piccinelli et al., 2024). Based on the visualization results in Fig. 6, compared to UniDepth and Metric3D, our method presents sharper details and more accurate structural relationships for the captured scenes.

Affine-invariant Depth Comparison. Though our method is trained for metric depth, we transform the predicted depth into affine-invariant depth for broader comparisons. As shown in Tab. 3, our model achieves performance comparable to methods specifically designed for affine-invariant depth, such as Marigold (Ke et al., 2024) and GeoWizard (Fu et al., 2024), despite being designed for metric depth. As show in Fig. 7, we provide visual results in both in-the-wild and synthetic scenarios. Our approach consistently demonstrates superior spatial structural understanding, such as accurately distinguishing the tree in the background or the Pisa tower, which is correctly inferred to be closer than the nearby church.

4.4 More 3D Vision Tasks

Monocular 3D Metrology. We evaluate the accuracy of our camera intrinsic estimation and metric depth prediction by estimating the true size of objects captured by cameras with varying focal lengths. For example, using the car shown in Fig. 8, we estimate the distance between the wheels. Compared to Metric3D(Yin et al., 2023), our method provides more accurate distance estimates across different

Figure 7: **Zero-shot qualitative affine-invariant depth results.** Our method demonstrates superior foreground-background differentiation (e.g., flower) and improved understanding (e.g., wall painting).

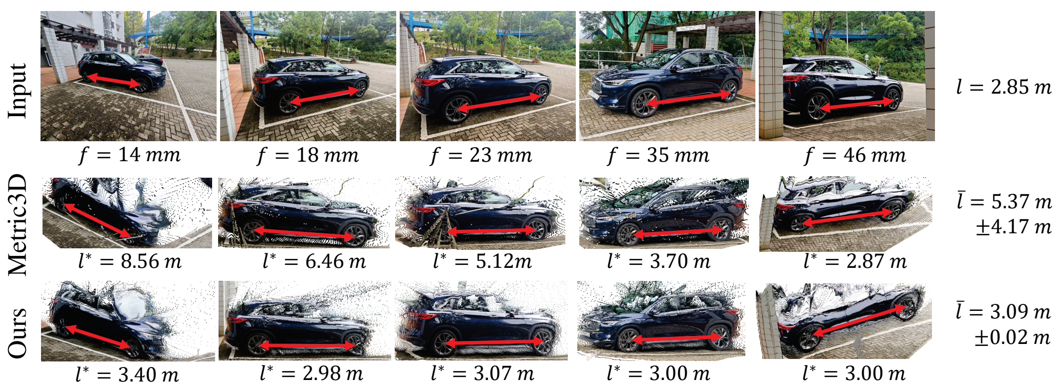


Figure 8: **Metrology of in-the-wild scenes.** Our method accurately recovers real-world metrics and demonstrates robustness to variations in focal length.

focal lengths and demonstrates robustness in both outdoor and indoor scenarios (see Fig. 11. However, performance decreases slightly with ultra-wide angles due to the scarcity of small focal length images in our training data, which could be improved by using more balanced datasets.

Sparse-view 3D Reconstruction & **Pose Estimation.** We further demonstrate that our estimated camera intrinsics can be effectively applied to sparse-view 3D reconstruction, especially when photos are taken with varying focal lengths. We evaluate the reconstruction results of Wang et al. (2024) with and without our estimated intrinsics. As shown in Fig. 9 and Fig. 13, reconstructions without intrinsic cues exhibit notable distortions and misalignments, whereas incorporating intrinsic cues significantly improves accuracy and alignment. Furthermore, as shown in Tab. 4, the reconstruction loss, represented by the mean relative distance between corresponding points, was reduced by around 20% on four in-the-wild scenes, confirming the effectiveness of using intrinsic cues for enhancing reconstruction quality. Additionally, as indicated in Tab. 5, pose estimation performance is also improved when intrinsic cues are used. Detailed experimental settings are provided in the appendix.

Table 4: **Relative distance error.** We compare the reconstruction performance with and without intrinsic cues.

	Sofa	Car	Pavilion	StoneWall
w/o. cue	1.67	0.87	1.03	1.43
w. cue	1.37	0.68	0.68	1.06

Table 5: **Pose error.** We compare the pose error with and without intrinsic cue.

	$t_{rel}\ (m)$	$r_{rel}\ (^\circ)$
w/o. cue	1.17	5.02
w. cue	0.63	2.30

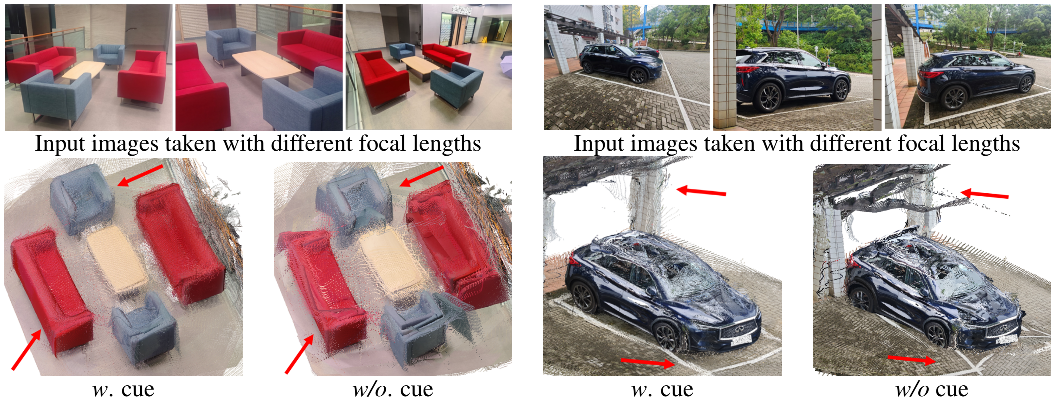


Figure 9: **Sparse View 3D Reconstruction with Intrinsic Cues.** We captured images with various focal lengths and present the reconstruction results. With intrinsic cues, our method achieves more accurate and better-aligned reconstructions.

4.5 Ablation Study

Ablation Study on Camera Calibration. We evaluate the effectiveness of our proposed camera image representation and multi-resolution noise strategy through an ablation study on the GSV dataset (Anguelov et al., 2010), which includes 20 different camera intrinsic parameters. The study, conducted over 15k iterations, focuses on field-of-view prediction errors. As shown in Tab. 6, the naive approach using $\boldsymbol{c} = [\theta, \phi, \theta]$ yields the poorest results due to the domain gap between the generated images and those produced by diffusion models. While multi-resolution noise(Kasiopy, 2023) improves performance slightly, it remains suboptimal. Incorporating our camera image representation significantly reduces the error, and the combination of both strategies produces the best results, demonstrating their complementary effectiveness.

Ablation on Metric Depth Estimation. We investigate the impact of our strategy on metric depth estimation by training a subset of our dataset for 20k iterations, with results shown in Tab. 7. Training solely on synthetic data has proven problematic, especially in indoor scenarios. Our second model, using a traditional multi-step denoising pipeline, integrates both virtual dense and sparse depth data but results in suboptimal performance due to the pipeline's inability to effectively recognize sparse areas in the ground truth. Additionally, prior methods that froze the VAE decoder during one-step training have shown to be inadequate for metric depth estimation, as demonstrated in our experiments. Omitting the camera image representation also slightly reduces accuracy. Interestingly, we found that the network can still produce relatively satisfactory results without explicit camera intrinsic guidance, contradicting previous studies that highlight the necessity of such information for metric depth estimation (Yin et al., 2023). We attribute this to the powerful pretrained SD model's capacity to capture subtle variations in camera intrinsic parameters.

Table 6: Ablation on Intrinsic Estimation.

Multi-Res. Noise	Camera Image	Mean Error (°)↓	Median Error (°) ↓
✗	✗	24.36	25.74
✓	✗	9.10	7.00
✗	✓	6.72	6.33
✓	✓	**2.54**	**2.01**

Table 7: Ablation on Metric Depth Estimation.

Ablation	NYU-v2			KITTI		
	$\delta_1 \uparrow$	$\mathrm{SI}_{\log} \downarrow$	A.Rel ↓	$\delta_1 \uparrow$	$\mathrm{SI}_{\log} \downarrow$	A.Rel ↓
Full Model	**85.8**	8.17	**13.5**	**89.1**	13.3	**11.7**
w.o Real data	26.5	8.80	39.8	69.6	19.0	24.7
w.o One step	77.1	11.9	17.1	76.7	17.1	18.9
w.o Decoder training	76.7	11.1	48.0	83.5	14.5	13.4
w.o Camera Image	83.8	**8.0**	14.6	87.8	**12.5**	12.1

5 Conclusion

In conclusion, we introduced *DM-Calib*, a diffusion-based framework leveraging our proposed Camera Image for monocular camera calibration. By utilizing the strong priors of stable diffusion models, *DM-Calib*effectively estimates camera intrinsic parameters. Extensive experiments demonstrate its superior performance across a range of 3D vision tasks, consistently outperforming baseline methods in real-world scenarios and varied imaging conditions. Future work could address ultra-wide-angle images by incorporating more diverse training data and improve inference efficiency by developing a few-step diffusion (Luo et al., 2023) model to further enhance 3D vision tasks.

REFERENCES

Adel Ahmadyan, Liangkai Zhang, Artsiom Ablavatski, Jianing Wei, and Matthias Grundmann. Objectron: A large scale dataset of object-centric videos in the wild with pose annotations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7822–7831, 2021.

Dragomir Anguelov, Carole Dulong, Daniel Filip, Christian Frueh, Stéphane Lafon, Richard Lyon, Abhijit Ogale, Luc Vincent, and Josh Weaver. Google street view: Capturing the world at street level. *Computer*, 43(6):32–38, 2010.

Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, et al. Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. *arXiv preprint arXiv:2111.08897*, 2021.

Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023.

Anand Bhattad, Daniel McKee, Derek Hoiem, and David Forsyth. Stylegan knows normal, depth, albedo, and more. *Advances in Neural Information Processing Systems*, 36, 2024.

Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2. *arXiv preprint arXiv:2001.10773*, 2020.

Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Xu Cao, Hiroaki Santo, Boxin Shi, Fumio Okura, and Yasuyuki Matsushita. Bilateral normal integration. In *European Conference on Computer Vision*, pp. 552–567. Springer, 2022.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021.

Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015.

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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

James M Coughlan and Alan L Yuille. Manhattan world: Compass direction from a single image by bayesian inference. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pp. 941–947. IEEE, 1999.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Xiao Fu, Wei Yin, Mu Hu, Kaixuan Wang, Yuexin Ma, Ping Tan, Shaojie Shen, Dahua Lin, and Xiaoxiao Long. Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image. In *ECCV*, 2024.

Simon Fuhrmann, Fabian Langguth, and Michael Goesele. Mve-a multi-view reconstruction environment. *GCH*, 3:4, 2014.

Guillermo Gallego, Elias Mueggler, and Peter F. Sturm. Translation of "zur ermittlung eines objektes aus zwei perspektiven mit innerer orientierung" by erwin kruppa (1913). *arXiv preprint arXiv:1801.01454*, 2018. URL https://arxiv.org/abs/1801.01454.

Gonzalo Martin Garcia, Karim Abou Zeid, Christian Schmidt, Daan de Geus, Alexander Hermans, and Bastian Leibe. Fine-tuning image-conditional diffusion models is easier than you think. *arXiv preprint arXiv:2409.11355*, 2024.

Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

Richard I Hartley. Euclidean reconstruction from uncalibrated views. In *Joint European-US workshop on applications of invariance in computer vision*, 1993.

Richard I. Hartley. Kruppa's equations derived from the fundamental matrix. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19:133–135, 1997.

Xiankang He, Guangkai Xu, Bo Zhang, Hao Chen, Ying Cui, and Dongyan Guo. Diffcalib: Reformulating monocular camera calibration as diffusion-based dense incident map generation. *arXiv preprint arXiv: 2405.15619*, 2024.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Masa Hu, Garrick Brazil, Nanxiang Li, Liu Ren, and Xiaoming Liu. Camera self-calibration using human faces. In *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*, pp. 1–8. IEEE, 2023.

Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu, Chunhua Shen, and Shaojie Shen. Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *arXiv preprint arXiv:2404.15506*, 2024a.

Wenbo Hu, Xiangjun Gao, Xiaoyu Li, Sijie Zhao, Xiaodong Cun, Yong Zhang, Long Quan, and Ying Shan. Depthcrafter: Generating consistent long depth sequences for open-world videos. *arXiv preprint arXiv:2409.02095*, 2024b.

Nianjuan Jiang, Ping Tan, and Loong-Fah Cheong. Symmetric architecture modeling with a single image. *ACM Transactions on Graphics*, pp. 1–8, 2009.

Linyi Jin, Jianming Zhang, Yannick Hold-Geoffroy, Oliver Wang, Kevin Blackburn-Matzen, Matthew Sticha, and David F Fouhey. Perspective fields for single image camera calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17307–17316, 2023.

Oğuzhan Fatih Kar, Teresa Yeo, Andrei Atanov, and Amir Zamir. 3d common corruptions and data augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18963–18974, 2022.

Kasiopy. Multi-resolution noise for diffusion model training. https://wandb.ai/johnowhitaker/multires_noise/reports/Multi-Resolution-Noise-for-Diffusion-Model-Training--VmlldzozNjYyOTU2?s=31, 2023. last accessed 17.11.2023.

Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Tobias Koch, Lukas Liebel, Marco Körner, and Friedrich Fraundorfer. Comparison of monocular depth estimation methods using geometrically relevant metrics on the IBims-1 dataset. *Computer Vision and Image Understanding (CVIU)*, 191:102877, 2020. doi: 10.1016/j.cviu.2019.102877.

Hyunjoon Lee, Eli Shechtman, Jue Wang, and Seungyong Lee. Automatic upright adjustment of photographs with robust camera calibration. *IEEE transactions on pattern analysis and machine intelligence*, 36(5):833–844, 2013.

Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9298–9309, 2023a.

Xian Liu, Jian Ren, Aliaksandr Siarohin, Ivan Skorokhodov, Yanyu Li, Dahua Lin, Xihui Liu, Ziwei Liu, and Sergey Tulyakov. Hyperhuman: Hyper-realistic human generation with latent structural diffusion. *arXiv preprint arXiv:2310.08579*, 2023b.

Yunxiang Liu and Ye Cui. Camera self-calibration network based on face shape estimation. In *2023 8th International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS)*, volume 8, pp. 144–150. IEEE, 2023.

Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9970–9980, 2024.

Bin Luo and Edwin R Hancock. Procrustes alignment with the em algorithm. In *International Conference on Computer Analysis of Images and Patterns*, pp. 623–631. Springer, 1999.

Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference, 2023.

Quang-Tuan Luong and Olivier D. Faugeras. Self-calibration of a moving camera from point correspondences and fundamental matrices. *International Journal on Computer Vision*, 22:261–289, 1997.

Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *The European Conference Computer Vision (ECCV)*, 2012.

Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. Unidepth: Universal monocular metric depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10106–10116, 2024.

Marc Pollefeys and Luc Van Gool. A stratified approach to metric self-calibration. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1997.

Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.

René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1623–1637, 2020.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 12179–12188, 2021.

Rene Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 44(03):1623–1637, 2022.

Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10912–10922, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 conference proceedings*, pp. 1–10, 2022a.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022b.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022.

Grant Schindler et al. An expectation maximization framework for simultaneous low-level edge grouping and camera calibration in complex man-made environments. In *Proceedings of CVPR*, volume 2, 2004.

Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Thomas Schöps, Johannes L. Schönberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

Steven M Seitz, Brian Curless, James Diebel, Daniel Scharstein, and Richard Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, volume 1, pp. 519–528. IEEE, 2006.

Gilles Simon, Antoine Fond, and Marie-Odile Berger. A-contrario horizon-first vanishing point detection using second-order grouping laws. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 318–333, 2018.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of rgb-d slam systems. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pp. 573–580. IEEE, 2012.

Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2446–2454, 2020a.

Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2446–2454, 2020b.

Bill Triggs. Autocalibration and the absolute quadric. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1997.

Igor Vasiljevic, Nicholas I. Kolkin, Shanyi Zhang, Ruotian Luo, Haochen Wang, Falcon Z. Dai, Andrea F. Daniele, Mohammadreza Mostajabi, Steven Basart, Matthew R. Walter, and Gregory Shakhnarovich. DIODE: A dense indoor and outdoor depth dataset. *CoRR*, abs/1908.00463, 2019. URL http://arxiv.org/abs/1908.00463.

Alexander Veicht, Paul-Edouard Sarlin, Philipp Lindenberger, and Marc Pollefeys. Geocalib: Learning single-image calibration with geometric optimization. *arXiv preprint arXiv:2409.06704*, 2024.

Alexander Veicht, Paul-Edouard Sarlin, Philipp Lindenberger, and Marc Pollefeys. Geocalib: Learning single-image calibration with geometric optimization. In *European Conference on Computer Vision*, pp. 1–20. Springer, 2025.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *CVPR*, 2024.

Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. Tartanair: A dataset to push the limits of visual slam. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4909–4916. IEEE, 2020.

Horst Wildenauer and Allan Hanbury. Robust camera self-calibration from monocular images of manhattan worlds. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2831–2838. IEEE, 2012.

Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, Deva Ramanan, Peter Carr, and James Hays. Argoverse 2: Next generation datasets for self-driving perception and forecasting. In *Advances in Neural Information Processing Systems*, 2021.

Alex Wong, Xiaohan Fei, Stephanie Tsuei, and Stefano Soatto. Unsupervised depth completion from visual inertial odometry. *IEEE Robotics and Automation Letters (RA-L)*, 5(2):1899–1906, 2020.

Jianxiong Xiao, Andrew Owens, and Antonio Torralba. Sun3d: A database of big spaces reconstructed using sfm and object labels. In *Proceedings of the IEEE international conference on computer vision*, pp. 1625–1632, 2013.

Guangkai Xu, Yongtao Ge, Mingyu Liu, Chengxiang Fan, Kangyang Xie, Zhiyue Zhao, Hao Chen, and Chunhua Shen. Diffusion models trained with large data are transferable visual models. *arXiv preprint arXiv:2403.06090*, 2024.

Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10371–10381, 2024.

Chongjie Ye, Lingteng Qiu, Xiaodong Gu, Qi Zuo, Yushuang Wu, Zilong Dong, Liefeng Bo, Yuliang Xiu, and Xiaoguang Han. Stablenormal: Reducing diffusion variance for stable and sharp normal. *arXiv preprint arXiv:2406.16864*, 2024.

Wei Yin, Xinlong Wang, Chunhua Shen, Yifan Liu, Zhi Tian, Songcen Xu, Changming Sun, and Dou Renyin. Diversedepth: Affine-invariant depth prediction using diverse data. *arXiv preprint arXiv:2002.00569*, 2020.

Wei Yin, Jianming Zhang, Oliver Wang, Simon Niklaus, Long Mai, Simon Chen, and Chunhua Shen. Learning to recover 3d scene shape from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 204–213, 2021.

Wei Yin, Jianming Zhang, Oliver Wang, Simon Niklaus, Simon Chen, Yifan Liu, and Chunhua Shen. Towards accurate reconstruction of 3d scene shape from a single monocular image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. Metric3d: Towards zero-shot metric 3d prediction from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9043–9053, 2023.

Xianggang Yu, Mutian Xu, Yidan Zhang, Haolin Liu, Chongjie Ye, Yushuang Wu, Zizheng Yan, Chenming Zhu, Zhangyang Xiong, Tianyou Liang, et al. Mvimgnet: A large-scale dataset of multi-view images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9150–9161, 2023.

Amir R Zamir, Alexander Sax, William B Shen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2018.

Cyril Zeller and Olivier Faugeras. *Camera Self-Calibration from Video Sequences: the Kruppa Equations Revisited*. PhD thesis, INRIA, 1996.

Menghua Zhai, Scott Workman, and Nathan Jacobs. Detecting vanishing points using global image context in a non-manhattan world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5657–5665, 2016.

Chi Zhang, Wei Yin, Billzb Wang, Gang Yu, Bin Fu, and Chunhua Shen. Hierarchical normalization for robust monocular depth estimation. *Advances in Neural Information Processing Systems*, 35: 14128–14139, 2022.

Hui Zhang, K Wong Kwan-yee, and Guoqiang Zhang. Camera calibration from images of spheres. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(3):499–502, 2007.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3836–3847, 2023.

Yueqiang Zhang, Langming Zhou, Haibo Liu, and Yang Shang. A flexible online camera calibration using line segments. *Journal of Sensors*, 2016(1):2802343, 2016.

Zhengyou Zhang. A flexible new technique for camera calibration. *IEEE Transactions on pattern analysis and machine intelligence*, 22(11):1330–1334, 2000.

Shengjie Zhu, Abhinav Kumar, Masa Hu, and Xiaoming Liu. Tame a wild camera: In-the-wild monocular camera calibration. In *NeurIPS*, 2023.

A More Implementation Details

A.1 Camera intrinsic prediction

We train our model on a diverse range of datasets, ensuring balance by selecting one dataset per batch with equal probability and sampling from it. Most datasets follow the setup of Zhu et al. (2023), with additional data incorporated to better leverage the capabilities of stable diffusion. A detailed description of the datasets is provided in Tab. 9. Notably, our training set includes more data compared to He et al. (2024). For a fair comparison, we also report our results using the same training dataset and results is shown in Tab. 8. Regarding the Camera Image, we normalize its values to the range $[-1, 1]$ by dividing by π, and instead of force-resizing, we pad the Camera Image to a resolution of 768×768. Unlike previous works (Zhu et al., 2023; He et al., 2024) that directly resize images to a fixed size, we resize the images while preserving their aspect ratios, padding the remaining areas with zeros. This approach is necessary because the data we used were collected with various aspect ratios even within a single dataset. Following the data augmentation strategy applied in (Zhu et al., 2023), we randomly scale images up to twice their original size and then crop them back to the original resolution, with the camera intrinsics adjusted accordingly.

Table 8: **Monocular Camera Calibration on Zero-Shot Datasets.** We report the calibration errors for both focal length and optical center. *Small* means we train our model with same dataset with Zhu et al. (2023) and He et al. (2024).

Method	Waymo		RGBD		ScanNet		MVS		Scenes11		Average	
	e_f	e_b	e_f	e_b	e_f	e_b	e_f	e_b	e_f	e_b	e_f	e_b
Ours-small	0.138	0.033	0.051	0.012	0.084	0.023	0.080	0.010	0.071	0.014	0.085	0.017
Ours	0.115	0.036	0.041	0.010	0.089	0.024	0.087	0.008	0.061	0.010	0.078	0.017

A.2 Metric depth prediction

For metric depth prediction, we do not pad the images. Instead, we resize the maximum dimension of the images to 768 while maintaining their aspect ratios. Additionally, we apply random horizontal flipping and random cropping to enhance dataset diversity even in one dataset. Inspired by (Fu et al., 2024), we incorporate a "scene distribution decoupler" into our model through text-guided conditioned depth generation. Specifically, we utilize the CLIP tokenizer and encoder to encode the terms "indoor geometry" and "outdoor geometry" for different environments. Based on this setting, we treat the metric depth with different scale factor for indoor and outdoor: $s = \{s_{\text{in}}, s_{\text{out}}\}$, and the depth label become $\boldsymbol{d}_s = \boldsymbol{d}/s_i$ with $s_i \in s$ to fit the output of the training VAE decoder.

A.3 More implementation details and discussions relating Figures and Tables.

Fig. 2: Our Camera Image is image-dependent, unlike other camera representations that are not. For other methods, lines can be plotted directly based on different FoV values. In contrast, we generate the line chart for the Camera Image using the GSV dataset (Anguelov et al., 2010), which includes 20 different types of cameras.

Fig. 9 & **Fig. 13**: We take 20 to 25 images with five different focal lengths (same image focal lengths as shown in Fig. 8) and perform the reconstruction based on these images. Surrouding are cropped for better visulization. Our method complements sparse-view reconstruction methods like Dust3r (Wang et al., 2024) by providing intrinsic information, rather than serving as a direct comparison. Dust3r (Wang et al., 2024) delivers less accurate intrinsic estimation because it focuses on sparse-view reconstruction by generating point clouds for image pairs and performing global alignment to jointly optimize intrinsic calibrations and poses. This process is less robust and often converges to a local minimum. In contrast, our method is specifically designed to recover camera intrinsics. The results demonstrate that Dust3r achieves more accurate reconstruction when equipped with our estimated intrinsics.

Tab. 5: The pose estimation is compared against pseudo-ground truth generated using COLMAP (Schönberger & Frahm, 2016) from 60 images of a single object, leveraging the ground truth focal length for improved accuracy. For the reconstruction, we select 20 of these images and compare the pose estimation with and without intrinsic cues. Note that SE(3) and scale alignment are applied for the comparison.

Table 9: **Datasets List for camera calibration.** List of the training and testing datasets: number of images, scene type, and method of calibration. SfM: Structure-from-Motion.

	Dataset	Images	Scene	Intrinsic
Training Set	NuScenes (Caesar et al., 2020)	28k	Outdoor	Calibrated
	KITTI (Cordts et al., 2016)	18 k	Outdoor	Calibrated
	CityScapes (Cordts et al., 2016)	23k	Outdoor	Calibrated
	NYUv2 (Nathan Silberman & Fergus, 2012)	6k	Indoor	Calibrated
	SUN3D (Xiao et al., 2013)	33k	Indoor	Calibrated
	ARKitScenes (Baruch et al., 2021)	48k	Indoor	Calibrated
	Objectron (Ahmadyan et al., 2021)	33k	Indoor	SfM
	MVImgNet (Yu et al., 2023)	27k	Indoor	SfM
	Hypersim (Roberts et al., 2021)	54k	Indoor	Synthetic
	Virtual KITTI (Cabon et al., 2020)	20k	Outdoor	Synthetic
	Taskonomy (Zamir et al., 2018)	420k	Indoor	Rendered
	TartanAir (Wang et al., 2020)	305k	Mix	Synthetic
Testing Set	Waymo (Sun et al., 2020a)	800	Outdoor	Calibrated
	RGBD (Sturm et al., 2012)	160	Indoor	Pre-defined
	ScanNet (Dai et al., 2017),	800	Indoor	Calibrated
	MVS (Fuhrmann et al., 2014)	132	Outdoor	Pre-defined
	Scenes11 (Chang et al., 2015)	256	Mixed	Pre-defined

Table 10: **Datasets List for Metric Depth estimation.** List of the training and testing datasets for metric depth estimation: number of images, scene type, and method of Acquisition.

	Dataset	Images	Scene	Acquisition
Training Set	Hypersim (Roberts et al., 2021)	54k	Indoor	Synthetic
	Virtual KITTI (Cabon et al., 2020)	20k	Outdoor	Synthetic
	Taskonomy (Zamir et al., 2018)	40M	Indoor	RGB-D
	TartanAir (Wang et al., 2020)	305k	Mix	Synthetic
	Argoverse2 Wilson et al. (2021)	403k	Outdoor	LiDAR
	Waymo Sun et al. (2020b)	223k	Outdoor	LiDAR
	Self-rendered	10k	Outdoor	Synthetic
Testing Set	Diode Vasiljevic et al. (2019)	771	Mix	LiDAR
	ETH3D Schöps et al. (2017)	454	Outdoor	RGB-D
	IBims-1 Koch et al. (2020)	100	Indoor	RGB-D
	NuScenes Caesar et al. (2020)	3k	Outdoor	LiDAR
	NYU Nathan Silberman & Fergus (2012)	654	Indoor	RGB-D
	VOID Wong et al. (2020)	800	Indoor	RGB-D

Tab. 3: We assess the generalization ability across five zero-shot datasets by aligning the predicted depth $\hat{\boldsymbol{d}}$ to the ground-truth depth $\boldsymbol{d}$ with a scale factor s and translation t, resulting in the aligned depth map $\boldsymbol{a} = s \times \hat{\boldsymbol{d}} + t$

Fig. 8 and Fig. 11: From a single input image, we first estimate the camera intrinsics and metric depth map, transform them into a 3D point cloud using the pinhole camera model, and calculate the 3D distance between key points.

Procrustes alignment: When pointcloud X is given, the relative pose can be obtained by Procrustes alignment (Luo & Hancock, 1999):

$$R^*, t^* = \underset{\sigma,R,t}{\arg\min} \sum_i \left\| \sigma(RX_i^{1,1} + t) - X_i^{1,2} \right\|^2,$$

where $X^{1,2}$ represents the pointmap of image 1 in the coordinate frame of image 2. Then, a global alignment of the pointmaps is performed to further refine the pose and obtain the final aligned pointcloud reconstruction

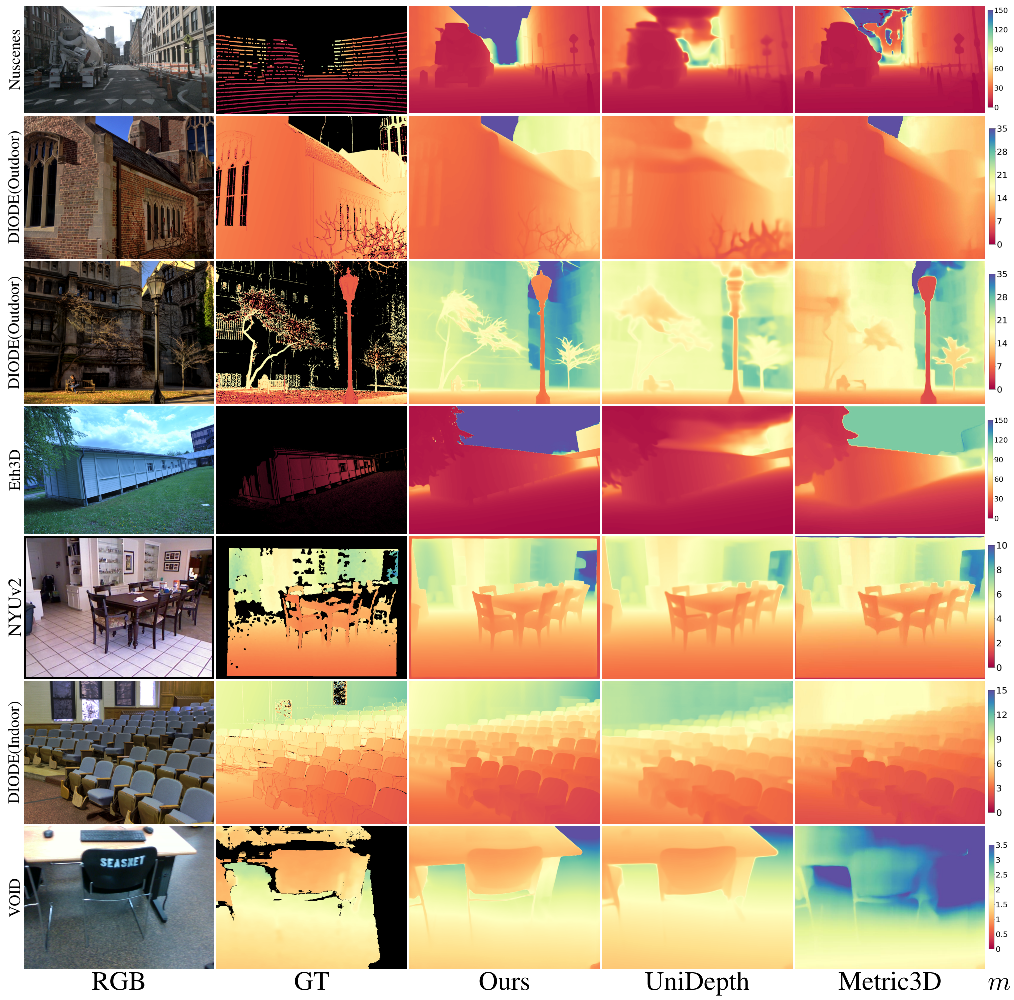


Figure 10: **Zero-Shot Metric Depth Estimation Results.** We present the predicted metric depth in both outdoor and indoor scenes. Our method provides more detailed results and recovers accurate metric depths.

B More experimental Results

B.1 Metric Depth

We show more qualitative metric depth prediction in Fig. 10.

B.2 Metrologie

We show more Metrologie results in Fig. 11 compared with Metric3D (Yin et al., 2023).

We also present the metrologie results for UniDepth (Piccinelli et al., 2024) in Fig. 12. While it shows some limitations in focal estimation, this leads to slightly less accurate visualizations.

B.3 3D reconstruction

We show more qualitative 3D reconstruction results in Fig. 13.

B.4 Mesh Reconstruction

By using our predict metric depth, we can deduce corresponding normal map, and mesh can be reconstructed via the depth and normal map using BiNI algorithm (Cao et al., 2022). We present the

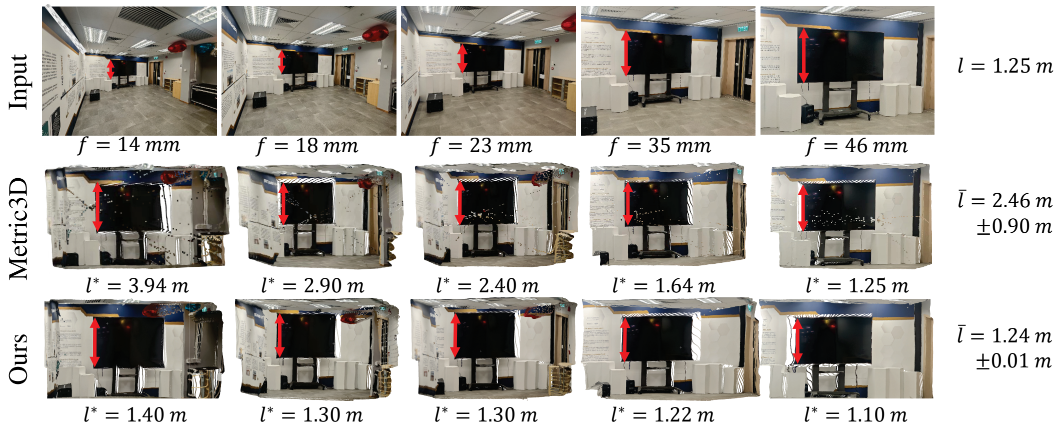


Figure 11: **Metrology of in-the-wild scenes.** Our method accurately recovers real-world metrics while demonstrating robustness to variations in focal length.

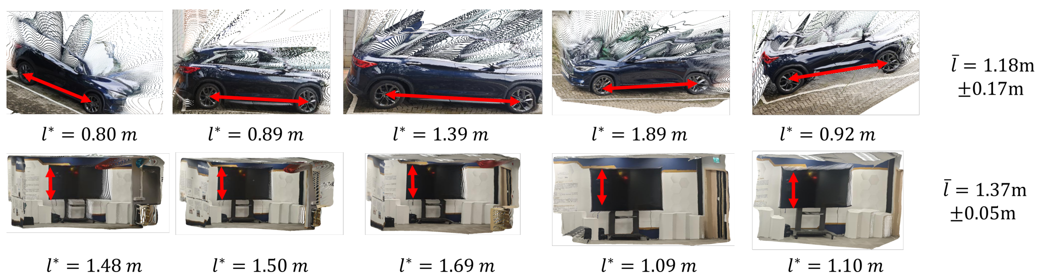


Figure 12: **Metrology of in-the-wild scenes for UniDepth.**

reconstruction result of Pisa tower in Fig. 7, and we show the reconstructed mesh in Fig. 14. Noting that we crop all background for better visualization.

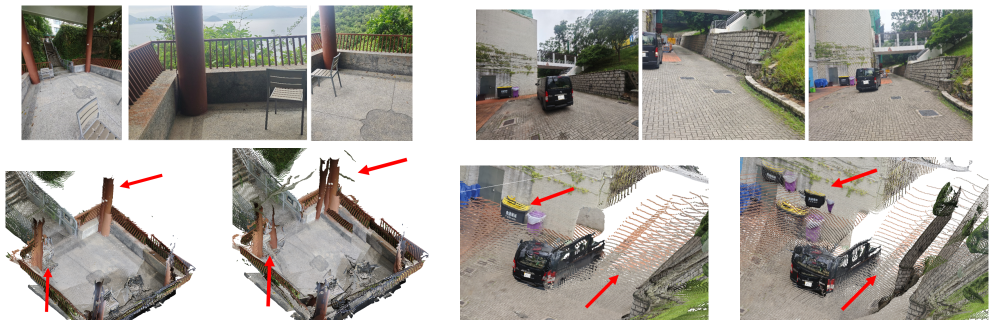
B.5 Single view 3D reconstuction

In this section, we present single-view 3D reconstruction of different camera focal length results using our estimated camera intrinsics and metric depth map. By applying the pinhole camera model, we transform the estimated intrinsics and depth map into a 3D point cloud. We demonstrate the robustness of our intrinsic estimation and depth prediction through in-the-wild single-view 3D reconstructions. Qualitative results can be found in Fig 15.

B.6 The Importance of Principal Point Evaluation and the Assessment of Both Vertical and Horizontal Focal Lengths

In our work, we evaluate the focal length as well as the principle points. Some previous works (Jin et al., 2023; Veicht et al., 2025) focuses solely on focal length. We prove the indispensability to evaluate the principal points. We have a significant amount of data where the principal point does not lie at the image center in certain datasets, and our model effectively learns the position of the principal points rather than ignoring them. To validate this, we conduct an ablation study comparing the error when assuming the principal point lies at the image center (e_b) with the error of our estimated principal point ($\hat{e}_b$). We show the results on Tab. 11.

Furthermore, not all datasets have $f_x = f_y$ (e.g., CityScapes dataset (Cordts et al., 2016) with $f_x = 2268.36$ and $f_y = 2225.54$). And our method is inherently capable of solving for both f_x and f_y and we take this into account to ensure more robust estimation and support future broader applications and datasets such as Diode (Vasiljevic et al., 2019).



W. Cue W.O Cue W. Cue W.O Cue

Figure 13: **Sparse view 3D reconstruction with intrinsic cue.** We captured images at different focal lengths and present the reconstruction results. With intrinsic cues, the reconstruction is more accurate and better aligned.

a). RGB b). Predicted Metric Depth c). Reconstructed mesh

Figure 14: **The reconstructed mesh using our predicted intrinsics and metric depth.**

B.7 THE IMPORTANCE OF CAMERA IMAGE IN METRIC DEPTH ESTIMATION.

The camera image (intrinsic information) is essential for robust and accurate metric depth estimation. We present the δ_1 results on three additional datasets in Tab.12, complementing the findings in Tab.7.

As shown, the absence of the camera image leads to a significant performance drop.

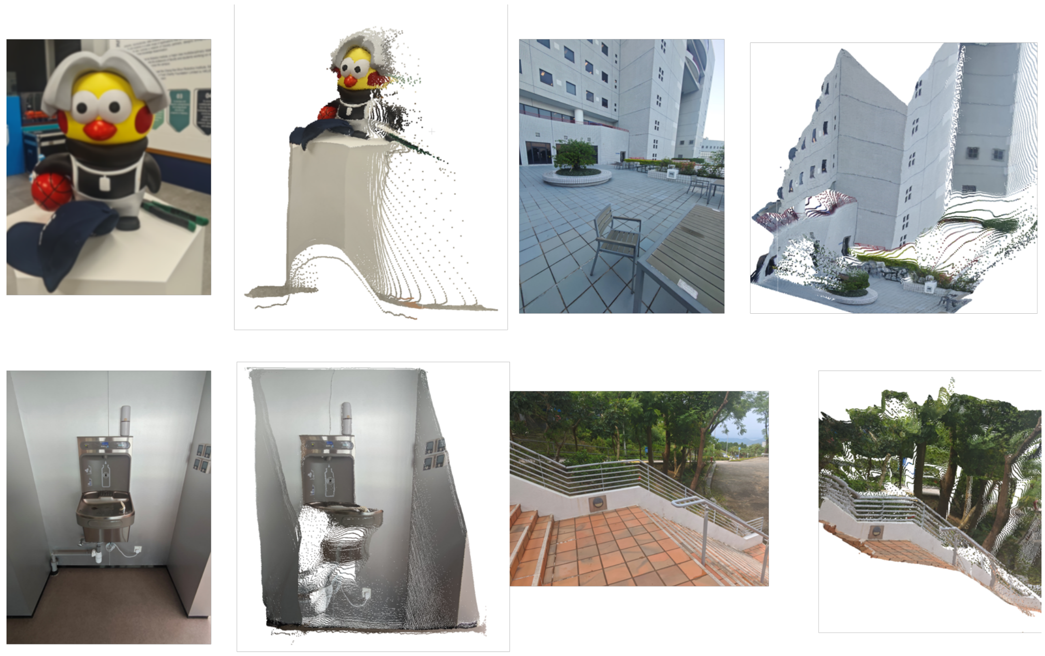
B.8 TEST-TIME ENSEMBLING

To reduce the stochasticity of the process, we aggregate five predicted camera images by taking their mean. This significantly minimizes the randomness of the diffusion model, as evidenced by the small standard deviation in Tab. 13.

Without the aggregation, the standard deviation is sometimes not negligible, as presented in Tab. 14.

Table 11: **Principal points error** We compare the error of principle point estimation when assuming principal point lies at the image center with the error of our estimated principal point.

	NuScenes	**KITTI**	**CityScapes**	**NYUv2**
e_b	0.051	0.021	0.055	0.050
$\hat{e}_b$	0.007	0.014	0.011	0.009



Input Images | Reconstructed 3D pointcloud | Input Images | Reconstructed 3D pointcloud

Figure 15: **The reconstructed pointcloud from images with different camera focal length using our predicted instrinc and metric depth.**

Table 12: Ablation study on the effectiveness of camera images for metric depth estimation.

	ibims	**Diode (indoor)**	**Diode (outdoor)**
w. camera image	88.7	50.1	41.0
w.o camera image	82.6	35.0	25.2

Table 13: **Standard Deviation of estimated intrinsics after ensembling.**

	Waymo	**RGBD**	**ScanNet**	**MVS**	**Scenes11**	**Average**
e_f	0.115 ± 0.008	0.041 ± 0.002	0.089 ± 0.002	0.087 ± 0.006	0.061 ± 0.006	0.078 ± 0.006
e_b	0.036 ± 0.001	0.010 ± 0.000	0.024 ± 0.000	0.008 ± 0.000	0.010 ± 0.001	0.017 ± 0.001

Table 14: **Standard Deviation of estimated intrinsics after ensembling.**

	Waymo	**RGBD**	**ScanNet**	**MVS**	**Scenes11**	**Average**
e_f	0.115 ± 0.035	0.041 ± 0.010	0.089 ± 0.024	0.087 ± 0.008	0.061 ± 0.009	0.078 ± 0.017
e_b	0.036 ± 0.012	0.010 ± 0.001	0.024 ± 0.001	0.008 ± 0.001	0.010 ± 0.001	0.017 ± 0.001