This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

AlignDiff: Learning Physically-Grounded Camera Alignment via Diffusion

Liuyue Xie[1*] Jiancong Guo[2] Ozan Cakmakci[2] Andre Araujo[3]
László A. Jeni[1] Zhiheng Jia[2]

[1] Carnegie Mellon University [2] Google [3] Google DeepMind

Abstract

Accurate camera calibration is a fundamental task for 3D perception, especially when dealing with real-world, in-the-wild environments where complex optical distortions are common. Existing methods often rely on pre-rectified images or calibration patterns, which limits their applicability and flexibility. In this work, we introduce a novel framework that addresses these challenges by jointly modeling camera intrinsic and extrinsic parameters using a generic ray camera model. Unlike previous approaches, AlignDiff shifts focus from semantic to geometric features, enabling more accurate modeling of local distortions. We propose AlignDiff, a diffusion model conditioned on geometric priors, enabling the simultaneous estimation of camera distortions and scene geometry. To enhance distortion prediction, we incorporate edge-aware attention, focusing the model on geometric features around image edges, rather than semantic content. Furthermore, to enhance generalizability to real-world captures, we incorporate a large database of ray-traced lenses containing over three thousand samples. This database characterizes the distortion inherent in a diverse variety of lens forms. Our experiments demonstrate that the proposed method significantly reduces the angular error of estimated ray bundles by $\sim 8.2^\circ$ and overall calibration accuracy, outperforming existing approaches on challenging, real-world datasets.

1. Introduction

Accurate camera calibration, involving the estimation of intrinsic parameters such as focal length and lens distortions and extrinsic parameters such as camera pose, is essential for robust 3D perception in real-world environments. However, prevailing methods typically address these components independently, focusing on either intrinsic calibration or pose estimation while relying on simplified camera models that are insufficient to capture complex real-world optical aberrations. This separation is fundamentally limiting, as intrinsic and extrinsic parameters are closely interdependent; inaccuracies in modeling lens distortions propagate directly to pose estimation errors. Consequently, joint optimization of intrinsic and extrinsic parameters is critical to ensure reliable calibration under unconstrained conditions.

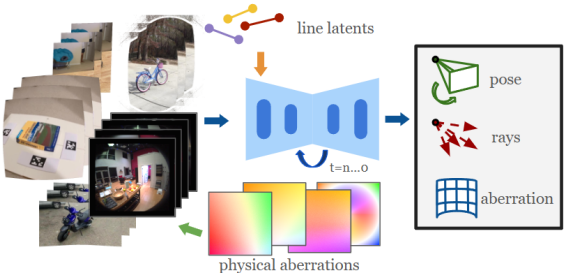

Figure 1. **AlignDiff** is proposed to address common image geometric aberrations with a unified ray camera representation while jointly recovering the camera extrinsics. With groundings on physical camera lens designs, as well as the disassociation of geometric cues from semantic features, it demonstrates an ability to generalize to real video sequences.

Recent diffusion-based and transformer-powered models, such as PoseDiffusion [55], RayDiffusion [63], and DiffusionSfM [1] have improved calibration accuracy and zero-shot generalizability by conditioning on image features extracted from vision encoders. However, these methods typically focus on high-level, semantic image features rather than on structural cues critical for modeling optical aberrations that directly impact geometry. This is particularly limiting because accurate intrinsic calibration is essential for precise extrinsic calibration: aberrations in the intrinsic parameters can significantly affect the accuracy of extrinsic parameters, making the two tightly interdependent.

We propose AlignDiff, shown Figure 1, a diffusion-based calibration framework, to learn the fine-grained ray profile from video sequences in world space, with a conditioning strategy that better captures fine-grained optical distortions. To shift focus from semantic to structural features, we condition the diffusion model with line embed-

*Work done during an internship at Google.

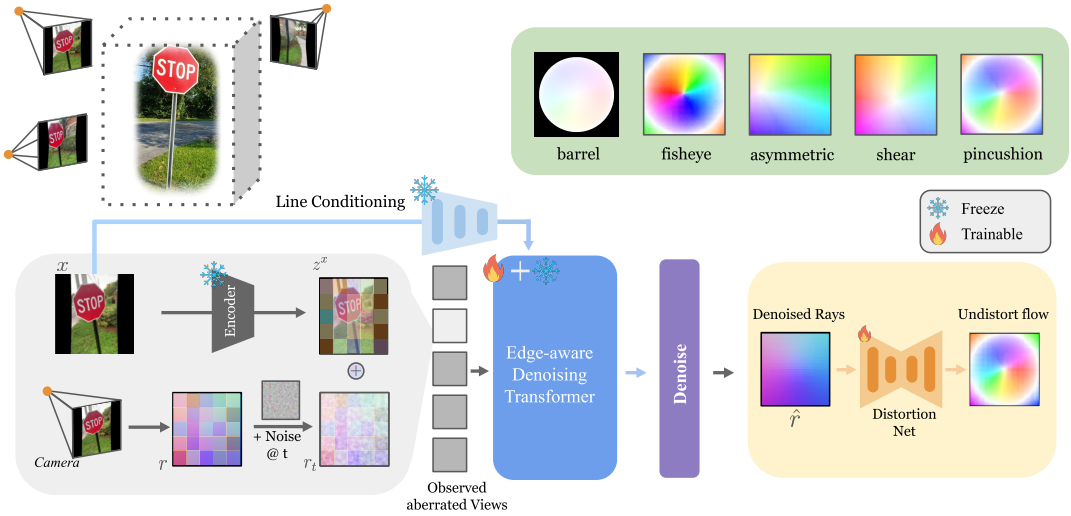

Figure 2. **AlignDiff Architecture.** We promote learning camera ray profiles in three main steps: geometric cue conditioning from line features, edge-aware attention, and physical camera groundings.

dings from a line detection network. This approach emphasizes the geometric structures, helping the model prioritize optical aberrations over object content. In addition, we further introduce edge-focused attention to further enhance the model's sensitivity to regions around edges, where aberrations often manifest most prominently.

An additional limitation of prior models is their reliance on simulated optical aberrations derived from predefined models, which often lack the subtlety of real-world aberrations, such as localized or asymmetric profiles. To address this, our framework incorporates real optical profiles from actual lens designs, grounding the model in authentic optical characteristics and boosting its generalizability.

In summary, our framework builds upon existing diffusion-based calibration techniques by incorporating targeted conditioning strategies and real-world optical data, providing a robust solution that accurately captures complex optical aberrations. By effectively addressing intrinsic aberrations, our model lays a solid foundation for accurate extrinsic calibration, making it highly applicable for in-the-wild scenarios requiring precise camera parameters.

- To our knowledge, AlignDiff is the first unified diffusion-based approach addressing complex optical aberrations in joint intrinsic and extrinsic calibration.
- By conditioning the diffusion model on line embeddings and incorporating edge-focused attention, our method prioritizes structural over semantic features, achieving generalization to diverse, real-world environments without requiring extensive retraining.
- Our approach incorporates authentic optical profiles from actual lens designs, grounding the model in real-world aberrations and improving generalization to natural images.

2. Related Works

Aberration representation for in-the-wild images. Traditional camera calibration methods model intrinsic characteristics, such as focal length and camera center, alongside distortion parameters that describe image geometric aberrations. Early deep learning-based methods [3, 9, 15, 45, 57, 59, 60] enabled calibration from images. Recent diffusion based approaches like PoseDiffusion [55], extend from prior approaches with improved zero-shot performance on wild captures. While effective for rectified images, these methods are limited in their ability to account for local optical aberrations resulting from factors like wear, temperature fluctuations, and lens designs.

Generic ray camera models were introduced to better capture local geometric distortions [43, 52] by modeling ray bundles across image patches, offering a more detailed representation of pixel deviations compared to the pinhole model. While accurate, they require denser calibration patterns, especially when optimized using methods like PnP [64]. Several recent work [15, 63, 66] adopted the ray representation in deep learning and proved its feasibility. Yet, these prior works remain largely restricted to the pinhole model and did not exploit the expressiveness of ray representation for aberration modeling. The recent blind camera undistortion line of work [29, 32] infers geometric displacements from a single image, often as a preprocessing step to provide rectified images.

Our work advances these approaches by introducing a unified ray-based framework that models both pinhole projections and local distortions. We directly estimate the ray profile and camera aberrations from a raw monocular sequence, enhancing generalizability in diverse real-world

conditions.

Joint Intrinsic and Extrinsic Calibration Classical camera calibration methods often assume a predefined camera model for each type of aberration, relying on calibration patterns and multi-view geometric constraints to estimate camera parameters from a sequence of images [13, 24–26, 33, 54, 55]. Early approaches solved joint intrinsic and extrinsic calibration through point matching across views [2, 3, 11, 22, 38, 49, 65], with refinements summarized in [31]. Structure-from-motion (SfM) techniques later optimized reprojection loss to refine parameters further.

For in-the-wild images, Hold-Geoffroy et al. [17] leveraged DenseNet [18] to estimate parameters such as horizon angle and vertical field of view. Other methods [20, 24] decouple parameter estimation from direct regression, and with integrated semantic or geometric guidance [6, 8, 21, 46]. Further advancements have shown that models pretrained with geometric understanding can act as better feature encoders [5, 15, 23, 40, 42, 45, 51, 53, 61]. Recent methods, including PoseDiffusion [55], DuSt3R [57], RayDiffusion [63] use diffusion and transformer to recover camera parameters or ray profiles directly from images, though they still assume fixed distortion models, limiting generalization to diverse real-world aberrations [31, 32]. The camera parameters are recovered by conditioning on local image features to capture local and cross-view geometric associations. However, these methods can yield residual errors around object contours due to reliance on semantic rather than purely geometric cues [10, 12, 27].

Our approach extends the ray representation without a fixed distortion model, disentangling semantic and geometric cues to reduce contour errors and enhance calibration accuracy. By integrating undistortion and calibration in a unified framework, we increase flexibility and applicability for real-world captures.

3. AlignDiff

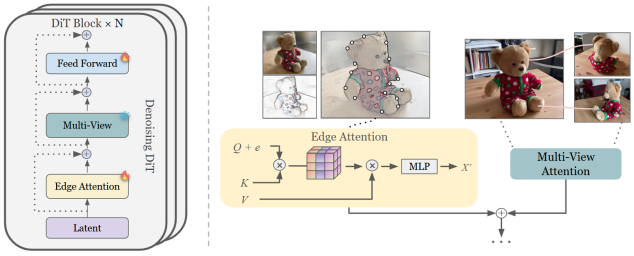

Figure 3. The latents are reweighted through edge-attention that aggregates a learned mask to the Query feature to promote information along edges where the geometric cues are more prominent. Following, the multi-view attention further captures features from different views.

We aim to recover generic cameras represented as ray bundles in world space from a set of N input images $\{\mathcal{I}_1, \mathcal{I}_2, \mathcal{I}_3, \ldots, \mathcal{I}_N\}$. Our approach, as presented in Figure 2, captures fine-grained geometric distortions while maintaining compatibility with classical parameterized camera models. This formulation is conceptually similar to works like [63, 66], but extends the ray camera model to account for local distortions (Sec. 3.1). In contrast, previous methods assume a pinhole camera model, neglecting distortions, even though they predict rays for local image patches. We adopt a Diffusion Transformer as the base architecture while enhancing this architecture by introducing geometric cues that guide ray directions based on local geometry, disentangling image perceptual features from distortions (Sec. 3.2). From the predicted rays, we propose DistortionNet to recover the underlying lens geometric aberration that describes the ray profile's deviation from a perspective pinhole camera with the same focal length and field of view (Sec. 3.3). We further discuss the selected optimization objectives in Sec. 3.4.

3.1. Generic Ray Camera Representation

The objective of camera calibration is to recover optimal parameters that describe the projection of 3D world points $\mathcal{X} \in \mathbb{R}^3$ to 2D sensor locations $\boldsymbol{x} \in \mathbb{R}^2$. Our method models this with a ray-based camera model, where each unit of ray $\boldsymbol{r}_i \in \mathbb{R}^6$ is stored on a grid centered on image patches. Rays originate from the camera center $\boldsymbol{r}_o \in \mathbb{R}^3$, with directions $\boldsymbol{r}_d \in \mathbb{R}^3$ derived by lifting 2D grid points to the camera frame using:

$$\boldsymbol{r_d} = \boldsymbol{K}^{-1}\mathcal{D}_\zeta(x), \tag{1}$$

where $\mathcal{D}_\zeta$ is the aberration function parameterized by ζ, and $\boldsymbol{K}$ is the intrinsic matrix, accounting for geometric aberrations. The ray bundle, transformed to world coordinates by rotation $\boldsymbol{R}$ and translation $\boldsymbol{t}$,

$$\boldsymbol{r}_w = \boldsymbol{R}^T\boldsymbol{r} + \boldsymbol{t}, \tag{2}$$

represents aberrated cameras in world coordinates, allowing holistic downstream recovery of rays, camera extrinsics, and optical aberrations.

Off-the-shelf image quality is completely characterized by the modulation transfer function (MTF) and distortion. The MTF performance includes all monochromatic and chromatic aberrations. Typically, the driving residual aberration in high-quality off-the-shelf lenses is optical distortion (W_{311} in the wavefront to third order[47]), therefore, in this paper, our ray trace models focus on extracting the optical distortion characteristics from our dataset of lenses.

3.2. Geometry-controlled ray diffuser

We train a multi-view diffusion model $\mathcal{M}_\phi$ that takes multiple images of a 3D scene as input and generates the corresponding output camera ray bundles given their geometric aberration cues. Specifically, given N conditional views

containing the images and their corresponding geometric aberration latents $\mathcal{Z}^g$ from a line-segment detection network, the model learns to capture the joint distribution of N target ray bundles $\mathcal{R}^{tgt} = \{\mathcal{R}_1, \mathcal{R}_2, \mathcal{R}_3, \ldots, \mathcal{R}_N\}^{tgt}$ from noised ray bundles $\mathcal{R}^{\epsilon}$:

$$p(\mathcal{R}^{tgt}|\mathcal{I}, \mathcal{Z}^g, \mathcal{R}^{\epsilon}). \tag{3}$$

Diffusion model architecture. Diffusion models approximate the data manifold by learning to invert a diffusion process from data to a presumed distribution through a denoising process. Our adopted Denoising Diffusion Probabilistic Model (DDPM) specifically defines the noise distribution to be Gaussian and transitions to the next noising step:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbb{I}), \tag{4}$$

where $\{\beta_1, \beta_2, ..., \beta_T\}$ are variances within the $\mathbf{T}$ step noising schedule, and $\mathbb{I}$ is the identity matrix. Similar to prior works [35, 55, 63], we adopt a DiT [16] as the backbone of the diffusion model, receiving image latent embeddings conditioned on aberration cues. Given a set of sequential images and target ray bundles, the model encodes each image into a latent representation $\mathcal{Z}_i^c$ through an image feature encoder TIPS [34], as well as a conditioning branch comprised of multiple ResNet [14] blocks. The image latent feature is concatenated to a noised ray bundle of the same dimension. Then, the diffusion model is trained to estimate the joint distribution of the latent-conditioned ray bundle given the conditioning geometric cues. We initialize the model from a DiT model trained for 3D shape generation, with an input resolution of $448 \times 448 \times 3$. We directly inflate the latent space of the original DiT to connect with the concatenated features of noised rays and image latents, while inheriting the rest of the trained model parameters to leverage the perceptual knowledge.

Utilizing geometric cues. The conditioning branch is designed to extract guidance information that describes geometric aberration cues and align it with the denoising features. The conditioning branch is a trained line segment detection network [28]. We observe that the controls maintain a high level of consistency with the denoising features and eliminate the need to be inserted at multiple stages, as also indicated in [41]. We thus integrate the controls at a single middle block into the denoising model by adding them to the denoising features after the cross-normalization. It can serve as a plug-and-play guidance module that provides geometric aberration information about the conditioning views, such that the diffuser model is aware of the levels and types of aberration from geometric cues.

Edge Attention. We introduce Edge Attention responsible for parsing features along the edges, as shown in Figure 3, which highlights the edge-aware attention mechanism designed to prioritize geometric cues along image edges. In particular, edge attention takes the input embedding of an image sequence $\mathcal{Z}^c \in \mathbb{R}^{N\times H\times W\times C}$, allowing it to perform self-attention operations across the edges. We apply a patchification operator with patch size $p \times p$ to generate patch tokens $\boldsymbol{t} \in \mathbb{R}^{L\times 3}$ with $L = (N \times H \times W/p^2)$ denoting the length of patchified tokens. To enable the model to disentangle the geometric information from the semantics, we integrate the edge information $e_{i\ldots N}$ for guidance into the self-attention mechanism:

$$Attention(Q, K, V) = Softmax(\frac{(Q + e^T)K}{\sqrt{H}}V) \tag{5}$$

where H is the dimension size of each head. The edge embedding is designed as a soft weighting to highlight the features along the geometrically prominent areas. The image edges are first extracted from the original image sequences using a Canny Edge detector. We then patchify them and apply a 3D convolution, mapping the embedding as $e_i \in \mathbb{R}^L$. Ultimately, we obtain the aggregated features $\hat{\mathcal{Z}}_i$, and we employ a feedforward network as in the original DiT before proceeding to the subsequent blocks.

3.3. Aberration profiling with DistortionNet

We adopt warping flow as a unified representation to model geometric aberrations of the image compared to ideal pinhole cameras, similar to prior blind camera undistortion frameworks [29, 32]. To query the aberration profile, we connect the denoised rays to a network devised to estimate deviations of predicted ray bundles to that of a pinhole camera. The DistortNet $\mathcal{M}_\theta$, with a backbone of MAE-Tiny [56], learns the mapping from the predicted set of ray bundles $\mathcal{R}^{tgt}$ to a set of warping flow maps $\mathcal{F}^{tgt}$. The learned flow maps can be applied to the image sequences for image undistortion.

3.4. Optimization objectives

We describe the optimization objectives for estimating ray bundles and extracting the distortion map from N images $\{\mathcal{I}_1, \mathcal{I}_2, \ldots, \mathcal{I}_N\}$. Estimating distorted ray bundles from in-the-wild images is under-constrained and affected by uncertainties from limited overlap and artifacts (e.g., motion blur, lighting changes). To address this, we train the network with a denoising process that reverses noised samples to match the reference:

$$\mathscr{L}_{denoise}(\phi) = \mathbb{E}_{t,\boldsymbol{r}_o,\epsilon}||\boldsymbol{r}_o - \mathcal{M}_\phi(\boldsymbol{r}_t, t)||^2, \tag{6}$$

where $\boldsymbol{r}_o$ is the original ray bundle, $\mathcal{M}_\phi$ is the denoising model, and $\boldsymbol{r}_t$ is the sample with Gaussian noise at time t:

$$\boldsymbol{r}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{r}_o + \sqrt{1-\bar{\alpha}_t}\epsilon, \tag{7}$$

where $\epsilon \sim \mathcal{N}(0, \mathbb{I})$. The denoising reverts to a regressive loss, allowing additional geometric constraints.

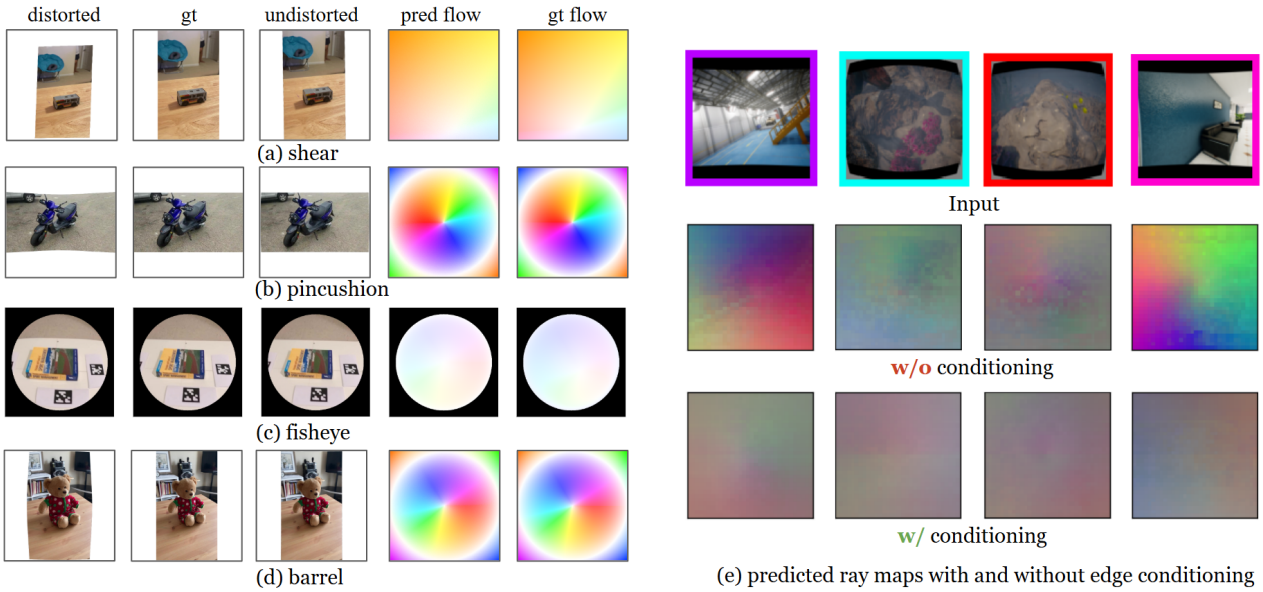

Figure 4. Recovered aberration profile and undistorted images. From denoised rays in world space, the DistortNet estimates the aberration pattern represented in warping flow. The undistorted images maintain coherent structural distribution compared to aberration-free images.

Recovering local camera ray profiles from an image sequence is susceptible to rotational ambiguities, leading to errors in ray bundle and pose estimates. We introduce a ray directional loss to measure angular errors in ray bundles:

$$\mathscr{L}_{angular}(\phi) = cos^{-1}\left(\frac{r_o \cdot \mathcal{M}_\phi(\boldsymbol{r}_t, t)}{|r_o||\mathcal{M}_\phi(\boldsymbol{r}_t, t)|}\right), \quad (8)$$

DistortNet is trained to regress backward flow, representing deviations from perspective ray bundles, by comparing each predicted flow map $\boldsymbol{f}_o$ to the reference set $\mathcal{F}^{tgt}$:

$$\mathscr{L}_{distort}(\theta) = \mathbb{E}_{t,\boldsymbol{f}_o,\boldsymbol{r}_t}||\boldsymbol{f}_o - \mathcal{M}_\theta(\mathcal{M}_\phi(\boldsymbol{r}_t, t))||^2. \quad (9)$$

This way, we recover the camera distortion maps at each image pixel, which can be used to directly unwarp the images and their associated rays.

3.5. Physical aberration groundings

We use the largest available lens database to ray trace and extract distortion characteristics from about 3000 optical systems. We wrote custom scripts that extracted each lens prescription from the LensView database and used commercial raytracing software to extract the optical distortion maps. These systems include camera lenses, lithography lenses, and freeform optics, as shown in Fig 5. These systems produce a diverse set of optical distortion functions that map the object to the image. Detailed description on the lens dataset can be found in Appendix A.

Since images within each dataset are usually captured with similar equipment sharing identical camera settings, to learn different camera modalities, we augment the sequences upon loading with sampled camera distortion profiles representing shear, barrel, fisheye, and pincushion. We further include a set of diverse optics dataset comprised of 3187 patented lens profiles from LensView [4]. These profiles span a wide range of professional and industrial lenses. Figure 5 shows examples from the grounding lenses, along with the distribution of the distortion with respect to the field of view (FOV), Numerical Aperture, and F-number. This augmentation introduces groundings in intrinsic and distortion profiles, addressing the scarcity of camera profiles within the datasets. We assess the out-of-distribution distortions not present in the lens set through experiments on the Aria Digital Twins dataset in Table 1.

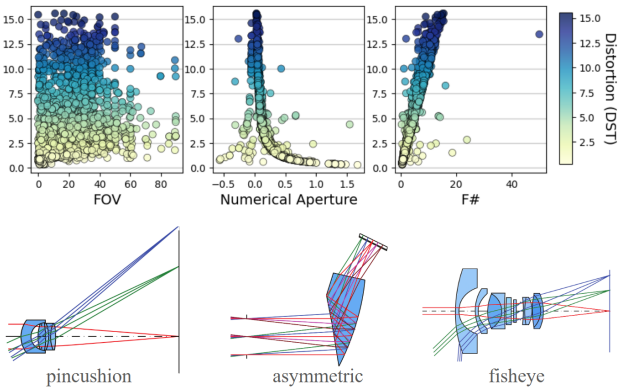

Figure 5. Ray-traced lens designs are utilized, enabling the accurate simulation of geometric aberrations. The aberrations are encoded as local geometric distortion [19], with a percentage deviation of pixels from their ideal positions on a regular grid.

4. Experiments

In this section, we first provide the implementation details of the proposed framework and then validate on two real world datasets with different camera modalities. Our method outperforms baseline approaches in both aberrated and non-aberrated images, achieving state-of-the-art results in quantitative and qualitative evaluations.

4.1. Implementation Details

We use a pre-trained TIPS [34] network as the image feature extractor, which gives high-quality, fine-grained image embeddings. We adopted a DiT [16] as the base structure of the denoising diffuser, and augmented the attention blocks with the proposed edge conditioning attention. We train our diffusion model with $T = 100$ timesteps, with the training taking roughly 3 days on 8 H100 GPUs.

Following prior works [33, 55, 63], we use the first camera as the coordinate anchor to define the scene. The scene is rescaled such that the first camera has a unit from translation and is rotated to grant the first camera identity rotation. In this way, all the following predicted cameras would use the first camera as a reference to express their relative orientations.

4.2. Dataset and Evaluation Metrics

Datasets. We chose CO3D [44] as the primary training dataset for the generation of ray maps. The dataset consists of roughly 37k 360° videos of common household objects from 51 MS-COCO categories. The dataset is annotated with COLMAP [49, 50] camera poses and intrinsics, with the frames undistorted with the estimated coefficients. Each video spans over on average 200 frames. The dataset provides a diverse profile of common objects sequences, making it suitable for training the network. In order to enhance the variety of training scenarios, we add MegaDepth [30] and TartanAir [58] as our additional training data.

Secondly, we evaluate the inference performance in Aria Digital Twins [39] that comprises 200 sequences and 400 min of videos that capture the daily activities of the common household using egocentric cameras. Its camera trajectories were provided by Optitrack [36] system and IMU sensor data on the Aria glass, further optimized by matching the two sources of estimated trajectories. The camera calibration parameters are provided by fitting a Kannala Brandt and Fisheye radial-tangential thin film model with the estimated trajectories. We use the same division as in [39], with a subset of 50 tests. With the two chosen datasets, we showcase the model's capabilities of estimating the ray maps from both common household videos and its ability to generalize to recent AR egocentric captures.

Evaluation protocol. We evaluated spare view predictions for recovered poses and the accuracy of the camera ray profile compared to the reference distorted camera, without employing the DistortionNet. In accordance with [63], we randomly sampled N images for an evaluation of N. Each reported accuracy is averaged over 5 runs to reduce stochasticity.

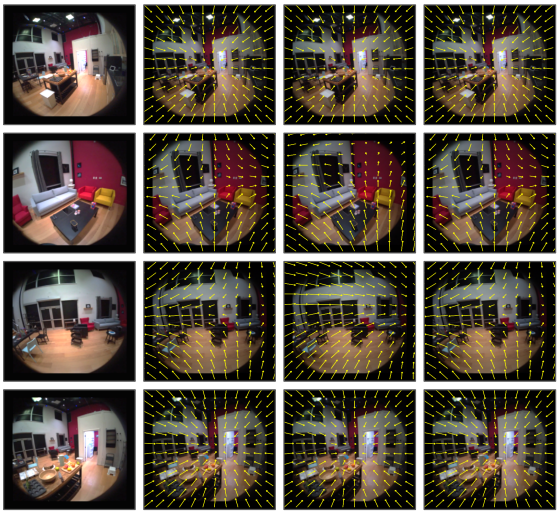

Figure 6. Generalization to out-of-distribution camera from Aria Digital Twins. Our approach produces better results compared to [63]. See supplementary for quantitative inference results.

Rotation accuracy. We first compute the relative rotations between each pair of cameras for both predicted and ground-truth poses. The errors are computed for the pairwise relative rotations and reported for the percentage of deviations less than 15°.

Camera center accuracy. Using the fitted ground truth and predicted poses, we align the cameras and compute the distance from the predicted camera center to the ground truth correspondence. The fraction of distances within 10% of the scene scale is reported.

Angular error. With the recovered ray bundles, we compute the distance of the predicted rays to the ground truth correspondences. The mean angular error in degrees is reported.

4.3. Evaluation

Baseline comparisons. In Table 1, we present the mean ray angular error measurements for the distorted CO3D dataset, demonstrating the effectiveness of our approach compared to existing methods. Our results indicate that our method excels at extracting precise, arbitrary ray profiles even in the absence of a prior camera model. This capability is particularly significant given the challenges of distorted sequences. Notably, using the same ray representation as RayDiffusion, our approach consistently surpasses its performance, where the diffusion model depends solely on image-based features. The experiments reveal that our method achieves a more robust disentanglement from local image content, effectively isolating structural cues from object-specific fea-

Table 1. **Quantitative evaluation of camera angular error** on CO3D and Aria Digital Twins datasets. Averaged angular error (degrees) across varying numbers of input images.

Method	2	3	4	5	6	7	8
CO3D – Seen Categories							
RayRegression [63]	18.4	20.6	21.7	21.9	20.3	23.0	23.7
RayDiffusion [63]	9.6	9.7	10.2	8.7	9.4	11.9	13.2
AlignDiff (Ours)	**3.4**	**3.1**	**2.8**	**3.2**	**3.0**	**3.6**	**4.1**
CO3D – Unseen Categories							
RayRegression [63]	24.1	22.5	21.6	24.8	23.9	25.1	26.2
RayDiffusion [63]	11.4	10.8	14.1	12.3	15.8	16.6	19.4
AlignDiff (Ours)	**4.6**	**5.2**	**5.8**	**5.7**	**6.4**	**6.8**	**8.2**
Aria Digital Twins							
RayRegression [63]	34.6	38.2	36.5	40.8	41.2	40.9	43.7
RayDiffusion [63]	28.4	29.1	34.7	37.2	40.4	42.3	45.6
AlignDiff (Ours)	**15.4**	**13.8**	**14.2**	**20.6**	**21.3**	**24.6**	**24.9**

Table 2. **Camera Rotation and Camera Center Accuracy** on Distorted CO3D dataset. Experiments using geometrically aberrated videos, comparing recent methods against AlignDiff.

	number of Images	2	3	4	5	6	7	8
	Rotation @ 15°							
Seen Cls	RelPose [62]	62.0	24.0	34.0	24.0	25.0	33.0	32.0
	RelPose++ [33]	72.5	73.4	73.6	74.7	75.0	76.0	75.7
	PoseDiffusion [55]	74.5	74.9	74.4	74.7	75.1	75.4	76.0
	RayDiffusion [63]	74.0	80.0	85.3	82.4	84.8	82.5	86.6
	AlignDiff (Ours)	**90.6**	**91.1**	**91.3**	**91.8**	**92.4**	**92.7**	**93.1**
Unseen Cls	RelPose [62]	61.0	27.0	37.0	26.0	30.0	26.0	25.0
	RelPose++ [33]	61.4	60.8	63.0	64.1	65.7	65.7	65.4
	PoseDiffsuion [55]	74.5	74.9	74.4	74.7	75.1	75.4	76.0
	RayDiffusion [63]	69.9	72.4	75.3	76.2	76.4	77.1	78.5
	AlignDiff (Ours)	**83.6**	**84.8**	**84.9**	**85.2**	**85.7**	**86.1**	**86.1**
	Camera Center @ 0.1							
Seen Cls	RelPose++ [33]	100	85.2	79.0	74.3	70.5	68.6	66.0
	PoseDiffusion [55]	100	72.3	56.7	53.6	52.4	57.1	52.1
	RayDiffusion [63]	100	73.4	72.5	71.1	71.0	70.6	70.2
	AlignDiff (Ours)	100	**92.5**	**90.7**	**88.6**	**87.2**	**85.8**	**84.6**
Unseen Cls	RelPose++ [33]	100	66.4	57.0	51.3	48.9	44.5	44.1
	PoseDiffusion [55]	100	72.3	56.7	53.6	52.4	57.1	52.1
	RayDiffusion [63]	100	74.8	71.6	70.2	67.3	66.7	62.5
	AlignDiff (Ours)	100	**86.3**	**84.2**	**80.5**	**75.2**	**73.0**	**70.1**

tures. This disentanglement minimizes residual errors tied to object appearance, allowing structural details to be more accurately emphasized.

Additionally, we report metrics for camera rotation accuracy and camera center accuracy, with aberrated images in Table 2 and rectified images in Table 3. Our approach consistently outperforms existing methods in estimating both camera position and orientation under aberrated conditions, confirming its resilience to challenging visual distortions. For rectified images, our conditioning method does not simply maintain performance but rather enhances it by recognizing and compensating for residual aberrations. This capability allows our approach to improve predictions even on images with minimal distortion, achieving a high level of precision and adaptability in handling diverse image qualities.

Table 3. **Camera Rotation and Camera Center Accuracy** on Rectified CO3D dataset. AlignDiff identifies subtle local distortions, achieving improvements over previous methods.

	number of Images	2	3	4	5	6	7	8
	Rotation @ 15°							
Seen Cls	COLMAP [48]	30.7	28.4	26.5	26.8	27.0	28.1	30.6
	RelPose [62]	56.0	56.5	57.0	57.2	57.2	57.3	57.2
	PoseDiffusion [55]	75.7	76.4	76.8	77.4	78.0	78.7	78.8
	RelPose++ [33]	81.8	82.8	84.1	84.7	84.9	85.3	85.5
	RayRegression [63]	88.8	88.7	88.7	89.0	89.4	89.3	89.2
	RayDiffusion [63]	91.8	**92.4**	92.6	92.9	93.1	93.3	93.3
	AlignDiff (Ours)	**92.6**	92.3	**92.7**	**93.0**	**93.6**	**93.4**	**93.5**
Unseen Cls	COLMAP [48]	34.5	31.8	31.0	31.7	32.7	35.0	38.5
	RelPose [62]	48.6	47.5	48.1	48.3	48.4	48.4	48.3
	PoseDiffusion [55]	63.2	64.2	64.2	65.7	66.2	67.0	67.7
	RelPose++ [33]	69.8	71.1	71.9	72.8	73.8	74.4	74.9
	RayRegression [63]	79.0	79.6	80.6	81.4	81.3	81.9	81.9
	RayDiffusion [63]	83.5	85.6	86.3	86.9	87.2	87.5	88.1
	AlignDiff (Ours)	**86.3**	**86.6**	**87.4**	**87.7**	**88.5**	**88.7**	**89.2**
	Camera Center @ 0.1							
Seen Cls	COLMAP [48]	100	34.5	23.8	18.9	15.6	14.5	15.0
	RelPose [62]	100	76.5	66.9	62.4	59.4	58.0	56.5
	PoseDiffusion [55]	100	77.5	69.7	65.9	63.7	62.8	61.9
	RelPose++ [33]	100	85.0	78.0	74.2	71.9	70.3	68.8
	RayRegression [63]	100	91.7	85.7	82.1	79.8	77.9	76.2
	RayDiffusion [63]	100	**94.2**	90.5	**87.8**	86.2	85.0	**84.1**
	AlignDiff (Ours)	100	93.7	**91.8**	87.5	**87.1**	**85.2**	83.9
Unseen Cls	COLMAP [48]	100	36.0	25.5	20.0	17.9	17.6	19.1
	PoseDiffusion [55]	100	63.6	50.5	45.7	43.0	41.2	39.9
	RelPose++ [33]	100	70.6	58.8	53.4	50.4	47.8	46.6
	RayRegression [63]	100	83.7	75.6	70.8	67.4	65.3	63.9
	RayDiffusion [63]	100	87.7	**81.1**	77.0	74.1	72.4	**71.4**
	AlignDiff (Ours)	100	**88.2**	81.0	**79.3**	**77.5**	**72.6**	71.0

Table 4. **Camera Rotation and Camera Center Accuracy** on Aria Digital Twins dataset.

	Number of Images	2	3	4	5	6	7	8
	Rotation @ 15°							
Unseen Cls	RelPose [62]	59.6	32.7	38.0	31.2	28.6	27.4	26.0
	RelPose++ [33]	62.8	61.3	62.6	62.8	63.2	64.9	64.8
	PoseDiffusion [55]	75.4	74.8	74.1	75.8	75.2	75.9	76.2
	RayDiffusion [63]	75.2	76.3	78.9	79.2	79.6	80.0	80.2
	AlignDiff (Ours)	**81.4**	**82.7**	**82.8**	**84.3**	**84.1**	**84.6**	**85.5**
	Camera Center @ 0.1							
Unseen Cls	RelPose++ [33]	100	70.3	56.1	52.8	48.6	46.4	44.7
	PoseDiffusion [55]	100	71.8	57.4	54.6	51.8	52.2	48.5
	RayDiffusion [63]	100	75.9	69.2	61.8	60.4	55.6	51.1
	AlignDiff (Ours)	**100**	**78.3**	**70.0**	**64.4**	**62.7**	**58.1**	**52.9**

Qualitative analysis of recovered cameras. We compare the generated cameras with the ground truth cameras in world space, as shown in Figure 7. The predicted ray moments and angular error from ground truth ray directions are also shown. Remarkably, the generated cameras exhibit little rotational difference from the reference cameras. While there are observed to have noticeable translation errors, we argue that this is due to scale ambiguities of the captures. The visualized ray moments present a smooth profile with very low angular errors. We further compare the Aria Digital Twins dataset for the inferred rays in vector form, shown in Figure 6. These further demonstrate that our approach can achieve robust and accurate ray profiles through camera

Table 5. Ablated framework design evaluated with averaged ray angular error.

Methods	2	3	4	5	6	7	8
Baseline	8.8	9.5	10.4	9.2	12.4	12.7	13.9
AlignDiff (w/o angular loss)	5.8	6.1	8.8	7.3	9.6	11.3	14.7
AlignDiff (w/o line conditioning)	6.7	8.4	9.8	9.9	10.3	10.1	12.8
AlignDiff (w/o edge-attentions)	5.3	7.1	7.4	9.2	10.8	11.3	10.5
AlignDiff (w/o optics grounding)	5.2	6.9	6.8	7.7	8.4	9.1	11.3
AlignDiff	**4.6**	**5.2**	**5.8**	**5.7**	**6.4**	**6.8**	**8.2**

Table 6. Evaluations on different image latent encoder choices.

Methods	2	3	4	5	6	7	8
DINOv2-s/14 [37]	8.2	8.1	8.6	10.4	9.8	10.1	12.5
DINOv2-g/14 [37]	**4.5**	6.4	6.3	6.6	**6.1**	7.8	9.6
TIPS-g/14 [34]	4.6	**5.2**	**5.8**	**5.7**	6.4	**6.8**	**8.2**

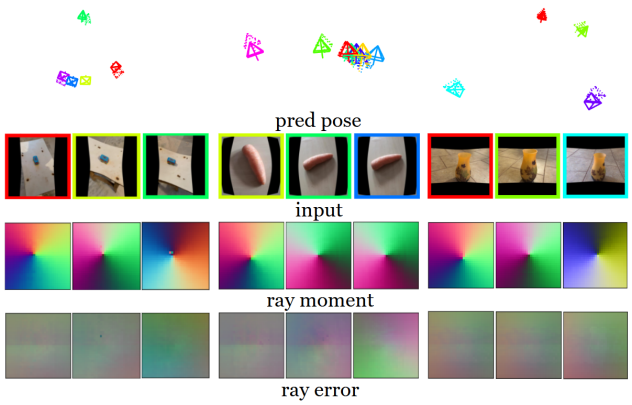

Figure 7. Recovered cameras from aberrated image sequences. Predicted and expected cameras are visualized in solid and dashed lines, respectively. The cameras align with the rotational orientation of the expected counterparts. The ray moment and residual errors show low levels of deviation from reference rays.

conditioning. Detailed quantitative evaluation on the Aria Digital Twins dataset are included in Appendix D.

Undistortion using predicted flow. We conduct image undistortion with the predicted aberration map, as shown in Figure 4. Since we adopt a unified representation for aberrations, the radial fisheye camera model's vignetting effect leads to mapping artifacts. The predicted aberration map overall correctly captures the image distortions and reliably remaps the images regardless of the aberration types.

Ablation study. We perform an ablation study on CO3D data to evaluate the impact of each conditioning module, enhanced geometric guidance, and additional losses, as shown in Table 5. For each comparison, we remove one conditioning module from the framework. The results show that all modules significantly improve performance, and latent conditioning of the line segment has the greatest effect. Figure 4 visualizes the ray residuals with and without aberration conditioning, further validating its effectiveness. Additionally, we test the angular loss, finding that our localized ray angular loss yields notable improvements.

While our multiview diffusion approach implicitly incorporates temporal consistency by explicit temporal losses, we chose to handle temporal consistency as a post-processing step to keep the model lightweight and directly applicable in real-world deployments. This approach reduces computational overhead while still providing meaningful consistency improvements, leaving more complex temporal modeling for future work. This is based on the premise that camera profiles remain generally consistent over short periods, barring changes due to factors like camera sensor temperature fluctuations. To test this, we implemented a post-processing step that rejects low-confidence pose estimations, which reduced the average ray angular error from 8.20 ° to 8.06 ° over 8 frames. Furthermore, enforcing temporal consistency on local camera ray profiles using reprojection mean squared error loss yielded an average reduction of 0.06° in angular error across the frames. We plan to integrate these enhancements, along with outlier rejection and a frame selection mechanism, into our future work.

Evaluations on image representation models. The image sequence encoder is designed to generate dense global embeddings that capture perceptual, geometric, and semantic representations. Among the available models, DINOv2 and TIPS are well suited for extracting high-quality dense features. We conduct an ablation study to evaluate these models' effectiveness in querying camera ray profiles. As shown in Table 6, the TIPS-g/14 model achieves the lowest angular errors on CO3D, with DINOv2-g/14 as a close competitor.

5. Conclusion

We introduce AlignDiff, a unified calibration framework for real-world optical aberrations, enabling the recovery of camera ray bundles, extrinsics, and aberration profiles from multi-view image sequences using simple conditioning techniques. First, we propose structural conditioning to separate image geometry from semantic features, promoting emphasis on structural cues for accurate ray profile estimation. Second, we aggregate features along edges, ensuring that high-quality image embeddings are paired with effective guidance to avoid semantic interference. Finally, we incorporate real-world lens profiles into training, grounding AlignDiff in actual camera designs. These contributions collectively enhance generalization and accuracy across diverse distortion profiles, as demonstrated by consistent improvements over strong baselines.

While AlignDiff performs well in controlled indoor settings, ray estimation may become unstable in arbitrary outdoor scenes due to scale ambiguity. To enhance robustness in such cases, integrating geometric priors like optical flow or combining traditional camera parameter estimation with network inference may be promising for future directions.

References

[1] Diffusionsfm: Predicting structure and motion via ray origin and endpoint diffusion, 2024. [Online; accessed 2025-03-06]. 1

[2] Tommaso Cavallari, Stuart Golodetz, Nicholas A. Lord, Julien Valentin, Victor A. Prisacariu, Luigi Di Stefano, and Philip H.S. Torr. Real-Time RGB-D Camera Pose Estimation in Novel Scenes Using a Relocalisation Cascade. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(10):2465–2477, 2020. 3

[3] Shuai Chen, Xinghui Li, Zirui Wang, and Victor Adrian Prisacariu. DFNet: Enhance Absolute Pose Regression with Direct Feature Matching. 2022. 2, 3

[4] Lambda Research Corp. Lensview optical design database. [Online; accessed 2025-03-07]. 5

[5] Andrei Cramariuc, Aleksandar Petrov, Rohit Suri, Mayank Mittal, Roland Siegwart, and Cesar Cadena. Learning camera miscalibration detection. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4997–5003, 2020. 3

[6] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperPoint: Self-Supervised Interest Point Detection and Description. 2017. 3

[7] Hao Feng, Wendi Wang, Jiajun Deng, Wengang Zhou, Li Li, and Houqiang Li. Simfir: A simple framework for fisheye image rectification with self-supervised representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12418–12427, 2023. 1

[8] Satoshi Fujimoto and Nobutomo Matsunaga. Deep feature-based rgb-d odometry using superpoint and superglue. *Procedia Computer Science*, 227:1127–1134, 2023. 3

[9] Fabrizio Gentile, Crescenzo Tortora, Giovanni Covone, Léon V. E. Koopmans, Rui Li, Laura Leuzzi, and Nicola R. Napolitano. LeMoN: Lens Modelling with Neural networks – I. Automated modelling of strong gravitational lenses with Bayesian Neural Networks. 2022. 2

[10] Behnam Gholami, Mostafa El-Khamy, and Kee-Bong Song. Latent feature disentanglement for visual domain generalization. *IEEE Transactions on Image Processing*, 2023. 3

[11] Diandian Guo, Deng-Ping Fan, Tongyu Lu, Christos Sakaridis, and Luc Van Gool. Vanishing-point-guided video semantic segmentation of driving scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3544–3553, 2024. 3

[12] Lanqing Guo, Renjie Wan, Wenhan Yang, Alex C Kot, and Bihan Wen. Cross-image disentanglement for low-light enhancement in real world. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(4):2550–2563, 2023. 3

[13] Annika Hagemann, Moritz Knorr, and Christoph Stiller. Deep geometry-aware camera self-calibration from video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3438–3448, 2023. 3

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 4

[15] 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. 2024. 2, 3

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. Technical report. 4, 6

[17] Yannick Hold-Geoffroy, Dominique Piché-Meunier, Kalyan Sunkavalli, Jean-Charles Bazin, François Rameau, and Jean-François Lalonde. A perceptual measure for deep single image camera and lens calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10603–10614, 2023. 3

[18] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269, 2017. 3

[19] ISO 17850:2015(en). Photography — Digital cameras — Geometric distortion (GD) measurements. Standard, International Organization for Standardization, 2015. 5

[20] Y. Y. Jau, R. Zhu, H. Su, and M. Chandraker. Deep keypoint-based camera pose estimation with geometric constraints. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4950–4957, 2020. 3

[21] Hanwen Jiang, Arjun Karpur, Bingyi Cao, Qixing Huang, and André Araujo. OmniGlue: Generalizable Feature Matching with Foundation Model Guidance. Technical report. 3

[22] Hanwen Jiang, Zhenyu Jiang, Kristen Grauman, and Yuke Zhu. Few-View Object Reconstruction with Unknown Categories and Camera Poses. 2022. 3

[23] Yang Jiao, Trac D. Tran, and Guangming Shi. Effiscene: Efficient per-pixel rigidity inference for unsupervised joint learning of optical flow, depth, camera pose and motion segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5538–5547, 2021. 3

[24] Linyi Jin, Jianming Zhang, Yannick Hold-Geoffroy, Oliver Wang, Kevin Matzen, Matthew Sticha, and David F. Fouhey. Perspective fields for single image camera calibration. In *CVPR*, 2023. 3

[25] Alex Kendall, Matthew Koichi Grimes, and Roberto Cipolla. Posenet: A convolutional network for real-time 6-dof camera relocalization. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2938–2946, 2015.

[26] Julius Kümmerle and Tilman Kühner. Unified intrinsic and extrinsic camera and lidar calibration under uncertainties. In *2020 IEEE International conference on Robotics and Automation (ICRA)*, pages 6028–6034. IEEE, 2020. 3

[27] Jinwoo Lee, Hyunsung Go, Hyunjoon Lee, Sunghyun Cho, Minhyuk Sung, and Junho Kim. Ctrl-c: Camera calibration transformer with line-classification. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16208–16217, 2021. 3

[28] Hao Li, Huai Yu, Jinwang Wang, Wen Yang, Lei Yu, and Sebastian Scherer. Ulsd: Unified line segment detection across pinhole, fisheye, and spherical cameras. *ISPRS Journal of Photogrammetry and Remote Sensing*, 178:187–202, 2021. 4

[29] Xiaoyu Li, Bo Zhang, Pedro V Sander, and Jing Liao. Blind geometric distortion correction on images through deep learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4855–4864, 2019. 2, 4

[30] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Computer Vision and Pattern Recognition (CVPR)*, 2018. 6

[31] Kang Liao, Lang Nie, Shujuan Huang, Chunyu Lin, Jing Zhang, Yao Zhao, Moncef Gabbouj, and Dacheng Tao. Deep learning for camera calibration and beyond: A survey. *arXiv preprint arXiv:2303.10559*, 2023. 3

[32] Kang Liao, Zongsheng Yue, Zhonghua Wu, and Chen Change Loy. Mowa: Multiple-in-one image warping model. *ArXiv*, abs/2404.10716, 2024. 2, 3, 4

[33] Amy Lin, Jason Y. Zhang, Deva Ramanan, and Shubham Tulsiani. RelPose++: Recovering 6D Poses from Sparse-view Observations. 2023. 3, 6, 7

[34] Kevis-Kokitsi Maninis, Kaifeng Chen, Soham Ghosh, Arjun Karpur, Koert Chen, Ye Xia, Bingyi Cao, Daniel Salz, Guangxing Han, Jan Dlabal, et al. Tips: Text-image pretraining with spatial awareness. *arXiv preprint arXiv:2410.16512*, 2024. 4, 6, 8

[35] Shentong Mo, Enze Xie, Ruihang Chu, Lanqing Hong, Matthias Niessner, and Zhenguo Li. Dit-3d: Exploring plain diffusion transformers for 3d shape generation. In *Advances in Neural Information Processing Systems*, pages 67960–67971. Curran Associates, Inc., 2023. 4

[36] Inc. DBA OptiTrack NaturalPoint. Optitrack system, 2024. 6

[37] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023. 8

[38] Linfei Pan, Marc Pollefeys, and Viktor Larsson. Camera pose estimation using implicit distortion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12819–12828, 2022. 3

[39] Xiaqing Pan, Nicholas Charron, Yongqian Yang, Scott Peters, Thomas Whelan, Chen Kong, Omkar Parkhi, Richard Newcombe, and Yuheng Carl Ren. Aria digital twin: A new benchmark dataset for egocentric 3d machine perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20133–20143, 2023. 6

[40] Chethan M. Parameshwara, Gokul Hari, Cornelia Fermüller, Nitin J. Sanket, and Yiannis Aloimonos. Diffposenet: Direct differentiable camera pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6845–6854, 2022. 3

[41] Bohao Peng, Jian Wang, Yuechen Zhang, Wenbo Li, Ming-Chang Yang, and Jiaya Jia. ControlNeXt: Powerful and Efficient Control for Image and Video Generation. 2024. 4

[42] Guilherme Potje, Felipe Cadar, Andre Araujo, Renato Martins, and Erickson R. Nascimento. XFeat: Accelerated Features for Lightweight Image Matching. 2024. 3

[43] Srikumar Ramalingam and Peter Sturm. A Unifying Model for Camera Calibration A Unifying Model for Camera Calibration. IEEE Transactions on Pattern Analysis and Machine Intelligence A Unifying Model for Camera Calibration. (7): 1309–1319, 2017. 2

[44] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common Objects in 3D: Large-Scale Learning and Evaluation of Real-life 3D Category Reconstruction. Technical report. 6

[45] Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From Coarse to Fine: Robust Hierarchical Localization at Large Scale. 2018. 2, 3

[46] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947, 2020. 3

[47] José Sasián. *Introduction to aberrations in optical imaging systems*. Cambridge University Press, 2012. 3

[48] Johannes L Schönberger and Jan-Michael Frahm. Structure-from-Motion Revisited. Technical report. 7

[49] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-Motion Revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3, 6

[50] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016. 6

[51] Cameron Smith, David Charatan, Ayush Tewari, and Vincent Sitzmann. Flowmap: High-quality camera poses, intrinsics, and depth via gradient descent. In *arXiv*, 2024. 3

[52] Peter Sturm and Srikumar Ramalingam. A Generic Concept for Camera Calibration. pages 1–13, 2004. 2

[53] Gabriele Trivigno, Carlo Masone, Barbara Caputo, and Torsten Sattler. The Unreasonable Effectiveness of Pre-Trained Features for Camera Pose Refinement. Technical report. 3

[54] Alexander Veicht, Paul-Edouard Sarlin, Philipp Lindenberger, and Marc Pollefeys. GeoCalib: Single-image Calibration with Geometric Optimization. In *ECCV*, 2024. 3

[55] Jianyuan Wang, Christian Rupprecht, and David Novotny. PoseDiffusion: Solving Pose Estimation via Diffusion-aided Bundle Adjustment. 2023. 1, 2, 3, 4, 6, 7

[56] Shaoru Wang, Jin Gao, Zeming Li, Xiaoqin Zhang, and Weiming Hu. A closer look at self-supervised lightweight vision transformers. In *Proceedings of the 40th International Conference on Machine Learning*. JMLR.org, 2023. 4

[57] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. DUSt3R: Geometric 3D Vision Made Easy. 2023. 2, 3

[58] 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. 2020. 6

[59] Scott Workman, Connor Greenwell, Menghua Zhai, Ryan Baltenberger, and Nathan Jacobs. Deepfocal: A method for direct focal length estimation. In *2015 IEEE International Conference on Image Processing (ICIP)*, pages 1369–1373, 2015. 2

[60] Shangrong Yang, Chunyu Lin, Kang Liao, and Yao Zhao. Innovating Real Fisheye Image Correction with Dual Diffusion Architecture. Technical report. 2

[61] Zhichao Yin and Jianping Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[62] Jason Y. Zhang, Deva Ramanan, and Shubham Tulsiani. Relpose: Predicting probabilistic relative rotation for single objects in the wild. page 592–611, Berlin, Heidelberg, 2022. Springer-Verlag. 7

[63] Jason Y. Zhang, Amy Lin, Moneish Kumar, Tzu-Hsuan Yang, Deva Ramanan, and Shubham Tulsiani. Cameras as Rays: Pose Estimation via Ray Diffusion. 2024. 1, 2, 3, 4, 6, 7

[64] Yinqiang Zheng, Yubin Kuang, Shigeki Sugimoto, Kalle Åström, and Masatoshi Okutomi. Revisiting the pnp problem: A fast, general and optimal solution. In *2013 IEEE International Conference on Computer Vision*, pages 2344–2351, 2013. 2

[65] Yichao Zhou, Haozhi Qi, Jingwei Huang, and Yi Ma. Neurvps: Neural vanishing point scanning via conic convolution. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2019. 3

[66] Shengjie Zhu, Abhinav Kumar, Masa Hu, and Xiaoming Liu. Tame a Wild Camera: In-the-Wild Monocular Camera Calibration. 2023. 2, 3