AlignDiff: Learning Physically-Grounded Camera Alignment via Diffusion

Supplementary Material

A. Grounding lenses

Lenses are designed with diverse characteristics, such as curvature, surface shapes, and defining parameters, to meet specific use cases. Compound lenses, widely adopted for objectives such as achieving predefined fields of view, enhanced resolution, or improved color intensity, consist of multiple surfaces with varying materials, coatings, and geometries. Commercially available lenses exhibit significant variation in design and aberration profiles, each tailored to distinct objectives. Common geometric profiles include barrel, pincushion, fisheye, symmetric, and asymmetric designs, as illustrated in Figure A2. Barrel and fisheye aberrations cause image points to appear closer to the center compared to a uniform reference grid, while pincushion aberrations push points outward relative to the grid.

To model lens profiles accurately, optical systems rely on ray tracing based on Snell's Law and paraxial optics. We show an example of a ray-traced optical system with geometric aberration in Figure A3. This approach derives spatially varying point spread functions (PSFs), relative illumination maps, distortion fields, and incident ray profiles. While spatially varying PSFs and illumination maps simulate perceptual aberrations on a scene image $\mathcal{I}s$, applying distortion fields alone offers a straightforward approach to simulate geometric aberrations. These aberrations are typically quantified as:

$$\mathcal{D}(\%) = \frac{d_{ad} - d_{ref}}{d_{ref}}, \tag{10}$$

where d_{ad} and d_{ref} denote the actual and reference distances from the image center, respectively, with d_{ref} derived via monochromatic paraxial ray tracing.

To compute distorted coordinates from an aberration profile, we perform bilinear interpolation on the distortion field at each pixel's location on the image plane.

B. Camera Uncertainty.

We present a visualization of aggregated results from our model across 20 runs for the same input sequence in Figure A4. Predicted poses are color-coded according to their respective input images. Sparse-view camera estimation inherently involves non-determinism due to scale ambiguity and symmetry. Despite these challenges, our method generates reasonable camera sets for each sequence. The predicted camera sets exhibit probabilistic variation, with larger variances observed at frames that have reflection-symmetric counterparts within the sequence.

Figure A1. **Aberration correction compared to existing method.** SimFIR [7] is a recent framework for blind image undistortion. Here we showcase the correction result from our approach and SimFIR.

C. Further details on baseline finetuning

In our comparisons, we chose PoseDiffusion, RelPose, RelPose++, RayDiffusion, and RayRegression as the data-driven baselines. The released checkpoints from these methods are not initially trained with geometrically aberrated images, leading to limited generalization for our proposed unified camera calibration task. We therefore finetune each model on the CO3D dataset using the same geometric aberration simulation as our framework for fair evaluations. For each of the model, we initialize the fine-tuning

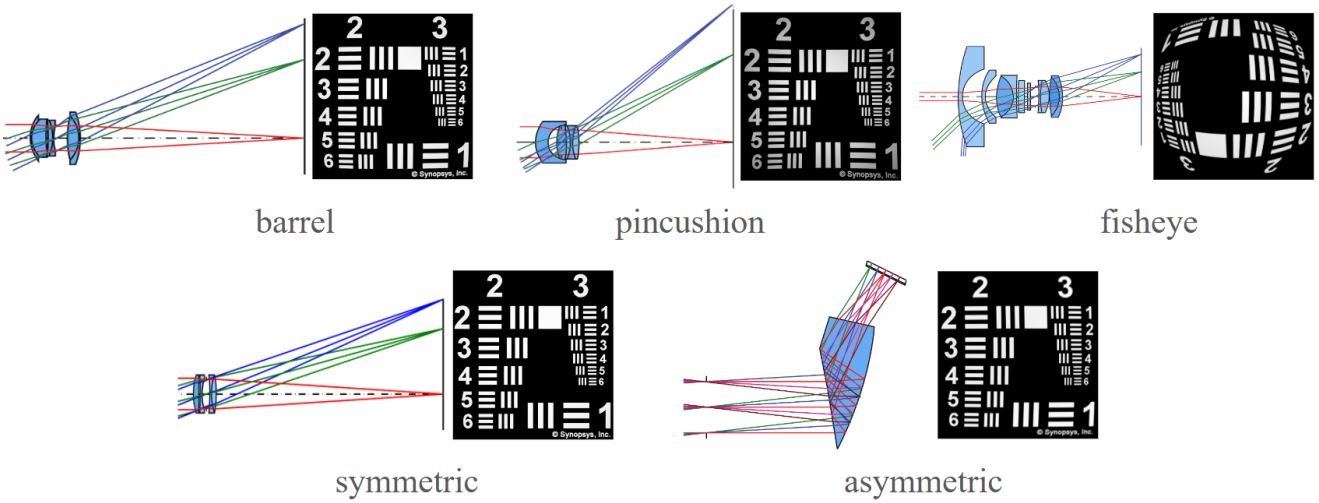

Figure A2. **Selected lens designs.** The lens dataset contains more than 3000 patented lens designs. The lens designs can be roughly categorized to resemble five aberration profiles. The geometric aberrations are converted into grid displacements, then applied to training sequences for simulation.

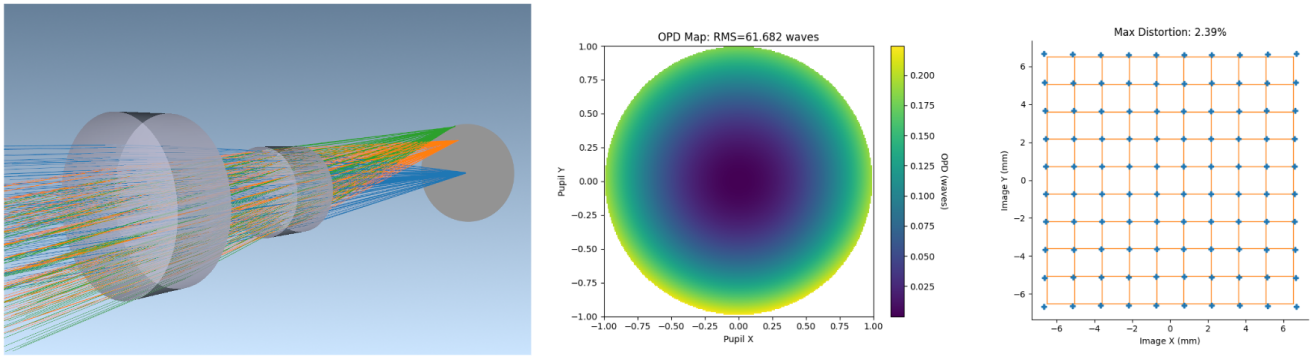

Figure A3. **Ray-traced lens.** For each lens design in the simulation data, we trace rays towards the image plane to derive the aberration representations. The reversed ray profiles originating from the image plane is to be estimated through our model. The pupil diagram and displacement grid are visualized, describing the ray-traced geometric aberrations.

with their respective publically released checkpoint. The dataloaders are modified to include the aberration simulations, where a random aberration is selected from the lens database, and then each image in the sequence goes through the same aberration as if each video is captured using the same camera model. We finetune for $10,000$ steps, and with a learning rate of 0.00005 for all models such that the losses are converged with the introduced camera modalities, without drifting far from the initial convergence state.

D. Aria Experiments

The Aria datasets are captured using geometrically aberrated fisheye cameras. Here, we test our framework for its reconstruction quality using the estimated camera pose and aberration profile.

Gaussian Splatting has emerged as a common 3D reconstruction framework, conditioned on properly undistorted sequential images and camera poses. It works as a solid testing ground to check the validity of our estimated cameras both for their spatial orientations and the aberration profiles. We attempted to reconstruct with 4 object-centric videos from the Aria Digital Twin Catalog dataset. The reconstruction results are shown in Figure A5, with corresponding videos showcased in the supplementary website.

E. Undistortion comparisons

Previous approaches have explored blind image undistortion from a single monocular image. While our method performs better when applied to a sequence of images, it can be adapted for single-image inference. We compare our

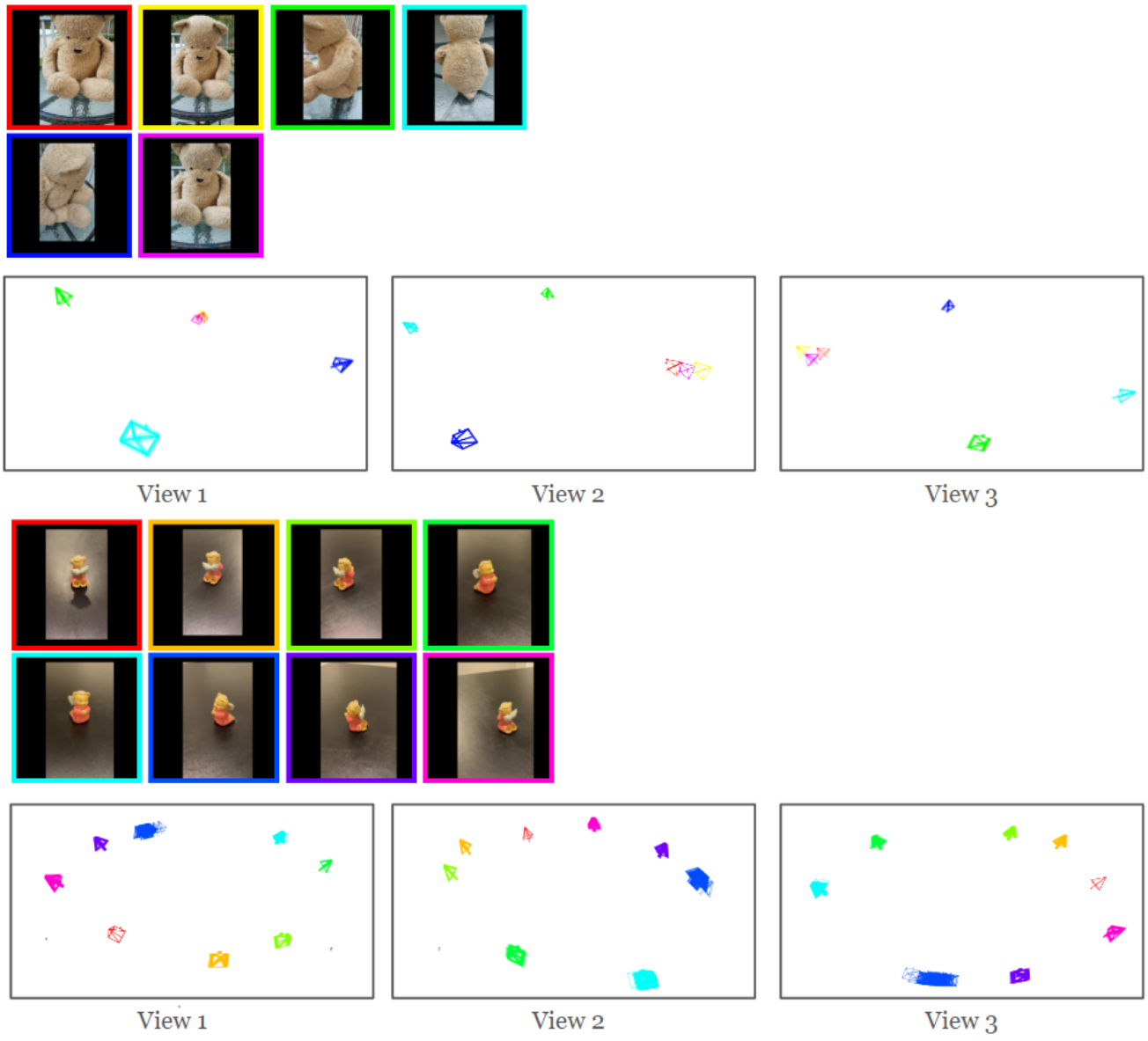

Figure A4. **Uncertainty Plot.** As a diffusion model, our framework makes prediction on the cameras in a probabilistic manner. The predictions are stochastic, with each set of predicted camera sequence resembling a valid ray bundle set.

undistortion results against the recent SimFIR method on the barrel, fisheye, and pincushion distortions, as illustrated in Figure A1. SimFIR employs a tailored representation for these distortions and predicts a vignetting mask for each image. Our method generally achieves improved undistortion quality, while the vignetting masks from SimFIR help mitigate visual artifacts near image boundaries.

Figure A5. **Gaussian Splatting Reconstruction.** Using the estimated ray bundles and distortion profiles, we process raw fisheye captures from the Aria Digital Twin Catelog dataset and train a Gaussian Splatting model for reconstruction.