

InFlux++: Real and Synthetic Data for Estimating Dynamic Camera Intrinsics

Erich Liang[✉], Caleb Kha-Uong[✉], Chinmaya Saran[✉], Sreemanti Dey[✉],
David W. Liu[✉], Junhan Ouyang[✉], Benjamin Zhou[✉], and Jia Deng[✉]

Princeton University, Princeton NJ 08544, USA

{erliang, ck2867, cs4046, sd9968, dl3533,
harryouyang, bz2883, jiadeng}@princeton.edu

Abstract. Camera intrinsics play a vital role in recovering 3D structure from 2D video. However, most 3D algorithms assume that intrinsics remain fixed throughout input video, an assumption that frequently fails for real-world in-the-wild videos. Consequently, estimating per-frame intrinsics from RGB images is critical for enabling 3D methods to operate robustly on dynamic intrinsics videos. Previously, InFlux has contributed to this research direction by establishing the first real-world benchmark with per-frame ground truth intrinsics for dynamic intrinsics videos, enabling systematic evaluation of algorithms for this task. Nevertheless, existing methods remain inaccurate due to two key obstacles: (i) training data for this task is scarce and lacks diversity in camera intrinsics; and (ii) existing benchmarks, including InFlux, are limited in scene and camera motion diversity, making it difficult to properly evaluate method performance. To address both gaps, we present InFlux++, consisting of two components. InFlux++ Synth is a large-scale procedurally generated synthetic video dataset with 441K+ annotated frames from 1841 high-resolution videos, providing accurate per-frame ground truth intrinsics for training dynamic intrinsics prediction models; a subset additionally includes per-frame camera pose, depth, and surface normals. The videos feature rich intrinsics diversity through changes in camera zoom and focus over time, as well as dynamic objects and realistic rendering effects such as lens distortion and defocus blur. InFlux++ Real is a large-scale real-world benchmark that extends InFlux with 514K+ newly captured frames across 334 high-resolution videos, spanning a wider range of scenes and camera motions. Finetuning existing intrinsics prediction methods on InFlux++ Synth consistently improves focal length estimation across both InFlux++ Real and InFlux, suggesting that synthetic supervision is a promising direction for RGB-based intrinsics prediction. For the dataset, benchmark, code, videos, submission instructions, and live leaderboard, please visit <https://influx.cs.princeton.edu/>.

Keywords: Dynamic Camera Intrinsics · Real-world and Synthetic Video · Procedural Generation

* Equal contribution.

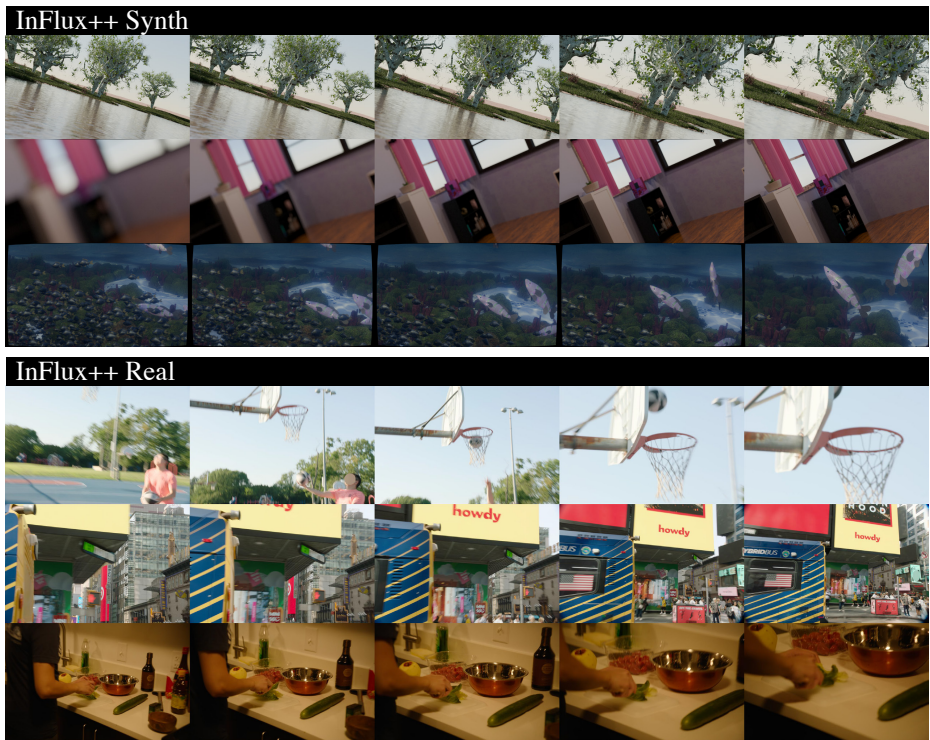


Fig. 1: A gallery of InFlux++, our data suite for dynamic intrinsics prediction. The top three rows show videos from InFlux++ Synth, a synthetic dataset with 441K+ frames across 1841 videos, featuring diverse intrinsics from changing zoom and focus, varying lens distortion, dynamic objects, and realistic optical effects such as defocus blur. The bottom three rows show videos from InFlux++ Real, a real-world dynamic intrinsics benchmark with 514K+ frames across 334 high-resolution videos. Compared to InFlux [12], it captures a wider range of indoor and outdoor scenes, natural daily activities, and richer camera motions with more translational motion.

1 Introduction

Camera intrinsics are fundamental to many real-world 3D systems, as they define the geometric mapping between 3D coordinates and the 2D image. Example applications include robotic depth perception and digital overlays in AR/VR.

Many modern 3D algorithms [8, 14, 15, 23, 26, 27] assume that camera intrinsics remain constant throughout video input, but real-world in-the-wild footage often violates this assumption. Intrinsics can vary during capture due to changes in zoom or focus distance, which alter the optical configuration of the lens.

As a result, accurately predicting per-frame camera intrinsics from RGB frames in dynamic intrinsics video is critical for extending standard 3D algorithms to real-world videos. Many 3D algorithms rely heavily on camera intrinsics for downstream geometric computations and assume that these parameters



Fig. 2: Gallery of InFlux++ Synth’s scene diversity. The dataset includes a wide range of procedurally generated indoor and nature environments with diverse lighting, materials, and scene composition. Indoor scenes feature varied room structures and non-Lambertian surfaces, while nature scenes span settings like forests, deserts, mountains, and underwater environments with rich terrain, vegetation, water bodies, and dynamic elements such as animals, falling leaves, and water motion.

remain constant throughout the input video. When intrinsics change during capture, this constant-intrinsics assumption leads to incorrect calibration values being used throughout the pipeline, degrading downstream computations. However, if intrinsics can be estimated accurately for each frame, they could be directly supplied to existing 3D algorithms, enabling reliable operation on dynamic intrinsics video without requiring fundamental architectural changes.

Previously, InFlux [12] has contributed to this direction of intrinsics prediction for dynamic intrinsics video by establishing the first real-world benchmark for this task, enabling systematic evaluation of prediction methods. However, state of the art (SOTA) methods such as [29] still incur large intrinsics prediction errors. When measuring prediction error via endpoint error (EPE), which quantifies the 2D distance between projections of the same 3D points under ground truth and predicted intrinsics, only 34.1% of projected points fall within 50 pixels of their ground truth locations. We identify two main obstacles in the field that limit intrinsics prediction performance.

The first obstacle to improving dynamic intrinsics prediction accuracy is the scarcity and limited diversity of training data for this task. Although many real-world datasets [3, 5, 24, 25] provide camera intrinsics, the intrinsics typically remain constant throughout each dataset. Panorama datasets [29, 30] can be used to generate image crops with custom focal lengths, but they cannot capture the realistic temporal changes in intrinsics or camera motion found in video, and their accuracy can be affected by manufacturing imperfections in 360° cameras or stitching artifacts introduced during panorama construction. Some synthetic datasets such as [31] provide videos with per-frame intrinsics, but most do not



Fig. 3: Gallery of InFlux++ Real’s scene diversity. Our benchmark contains a wide range of environments, activities, and viewpoints.

vary the intrinsics across frames or videos. Among the few that do, [28] varies focal length but does not model lens distortion or defocus blur, limiting its usefulness for training models for dynamic intrinsics video prediction.

The second obstacle to improving dynamic intrinsics prediction accuracy is the limited scene and camera motion diversity in existing benchmarks for this task. In terms of scene diversity, InFlux [12] outdoor coverage consists mostly of shots of outdoor university campus areas or natural landmarks, with relatively few shots in busy urban environments. Indoor coverage is limited to office or workspace settings, or empty laboratory spaces featuring staged skits, which do not capture the range of natural daily activities found in real-world footage. Regarding camera motion, InFlux predominantly features rotational motion, with little camera translation or parallax. This lack of varied camera motion may not adequately reflect the camera movement found in real-world videos.

To address these limitations, we introduce InFlux++, a comprehensive data suite for dynamic intrinsics prediction, comprising a large-scale synthetic training set and an extended real-world benchmark. InFlux++ Synth is a procedurally generated synthetic video dataset with 441K+ annotated frames from 1841 high-resolution videos, designed for supervised training of dynamic intrinsics prediction models. Building on Infinigen [19,20], we extend the renderer to generate accurate per-frame ground truth intrinsics with high diversity in camera parameters, achieved by varying both zoom and focus throughout each video. A subset of the dataset also provides camera pose, depth, and surface normals. The videos include dynamic objects, realistic optical effects such as lens distortion and defocus blur, and capture dynamic video sequences rather than isolated images, addressing the limited intrinsics diversity and lack of frame-to-frame variation

in prior datasets. InFlux++ Real extends the original InFlux benchmark with 514K+ newly captured frames across 334 high-resolution videos, directly addressing prior limitations in scene and motion diversity. Compared to InFlux, it includes substantially more outdoor urban environments, a broader range of indoor settings beyond offices and laboratory scenes, and a greater emphasis on naturally occurring real-world activities. It also features more camera translation than InFlux, producing greater parallax and better reflecting real-world camera motion. Together, these resources provide both the training diversity and benchmark coverage needed to systematically develop and evaluate robust dynamic intrinsics prediction methods. A gallery of InFlux++ can be found in Fig. 1.

Using InFlux++ Real and InFlux as evaluation data, we assess the performance of existing intrinsics prediction baselines [6, 7, 17, 23, 29, 30, 32]. We further finetune AnyCalib [29], the current SOTA method for dynamic intrinsics prediction, on InFlux++ Synth. Evaluation results show that while dynamic intrinsics prediction remains challenging, finetuning on InFlux++ Synth yields improved focal length estimation across both InFlux++ Real and InFlux. These results suggest that synthetic supervision is a promising direction for RGB-based dynamic intrinsics prediction. In summary, our contributions are:

- We introduce InFlux++ Synth, a procedurally generated synthetic video dataset with accurate and dynamic intrinsics spanning 441K+ frames across 1841 videos. A subset also includes pose, depth, and normals annotations. The videos feature diverse intrinsics from changes in zoom and focus, varying lens distortion, dynamic objects, and realistic optical effects like defocus blur.
- We introduce InFlux++ Real, a real-world dynamic intrinsics benchmark extending InFlux with 514K+ frames across 334 high-resolution videos. Compared to InFlux, it includes substantially more outdoor urban environments, a broader range of indoor settings, more naturally occurring real-world activities, and richer camera trajectories featuring more translational motion.
- We demonstrate that finetuning the SOTA method AnyCalib [29] on InFlux++ Synth leads to improved focal length estimation on InFlux++ Real and InFlux, suggesting that synthetic supervision is a promising direction for RGB-based dynamic intrinsics prediction.

2 Preliminaries

In this paper, we adopt similar camera and lens terminology established in [12]:

Camera Focal Length (CFL) is the distance from a camera’s optical center to its imaging plane along the optical axis. CFL corresponds to the f_x and f_y entries of the camera intrinsics matrix when measured in pixels.

Lens Focal Length (LFL) is the CFL of a camera when it is focused at infinity [21]. Effective CFL often differs from the LFL for real-world images.

Lens to Object Distance (LTO) is the distance from the camera’s optical center to the object in focus, measured along the optical axis.

Focus Distance (FD) is the distance along the optical axis from the camera’s sensor plane to the object in focus. It is the sum of CFL and LTO.

Zoom Lenses can produce dynamic camera intrinsics during video filming via mechanical shifts in the zoom ring, which controls LFL, and the focus ring, which controls FD. Some specialized zoom lenses report per-frame LFL and FD as lens metadata, which together fully characterize the optical state of the lens [1].

3 Related Work

3.1 Training Data for Dynamic Intrinsics Prediction

Real-World Training Datasets. Popular datasets like [3, 5, 25] provide camera intrinsics, but are captured with pre-calibrated camera rigs with fixed intrinsics, making them unsuitable for supervising dynamic intrinsics prediction. [24] records videos using a small set of different lenses. Although this introduces variation in camera intrinsics, the variation occurs only across different videos—intrinsics still remain fixed within each individual video sequence, making the dataset unsuitable for supervising dynamic per-frame intrinsics prediction. [11] provides COLMAP-derived intrinsics for Internet photo collections [23], but lacks temporal video structure and relies on reconstruction-derived pseudo ground truth. [9, 22] include time-varying intrinsics for AR device images, but the changes are incidental and small, arising from effects like autofocus or changes in device temperature. Panorama-based datasets such as [29, 30] construct images with varying intrinsics by cropping 360° panoramas. While this introduces large changes in intrinsics, the images lack temporal correspondence and rely on ideal panorama geometry, making them susceptible to stitching artifacts and lens imperfections. In contrast, InFlux++ Synth provides high-resolution synthetic video sequences with changing intrinsics and perfectly accurate per-frame annotations, enabling effective supervision for dynamic intrinsics prediction.

Synthetic Training Datasets. Synthetic datasets inherently provide perfectly accurate intrinsics, thanks to privileged knowledge of the rendering camera, but few are suitable for supervising dynamic intrinsics prediction. [10, 13, 31] use a fixed focal length for all renders, so they cannot be effectively used for training dynamic intrinsics prediction models. [28] provides videos with varying focal length, but it only uses the simple pinhole camera model for rendering. As a result, its videos lack defocus blur, which occurs in real camera systems and could serve as a useful training signal for intrinsics prediction. Its videos also lack lens distortion, which reduces the diversity of intrinsics. In contrast, InFlux++ Synth uses the thin lens model for video rendering, enabling realistic rendering effects like defocus blur. It also includes a data loader that applies customizable Brown-

Conrady [2] image distortion, thus providing diverse yet accurate intrinsics for realistic videos to support training of dynamic intrinsics prediction models.

3.2 Benchmarks for Dynamic Intrinsics Prediction

InFlux [12] is the closest existing benchmark, offering per-frame ground truth intrinsics across varied cameras and scenes. However, its scene coverage is narrow, skewing toward university campuses and natural landmarks for outdoor scenes, and office or laboratory spaces with staged subjects for indoor scenes. InFlux’s camera motion is also predominantly rotational, with little translation or parallax. In contrast, our benchmark InFlux++ Real covers a broader range of environments, including busy urban settings, everyday indoor activities like cooking and laundry, and diverse outdoor scenes such as sports and recreation. InFlux++ Real features longer videos with subjects performing natural everyday activities such as household chores or playing basketball. Camera motion includes more translational movement—including vehicle-mounted capture—resulting in richer parallax and a more faithful reflection of the diversity of real-world videos.

4 Methodology

In this section, we describe the construction of InFlux++, which consists of two parts: InFlux++ Synth, a large-scale synthetic video dataset with accurate per-frame intrinsics, realistic optical effects, dynamic objects, and pose, depth, and normals for a subset of frames; and InFlux++ Real, a real-world benchmark that extends InFlux [12] with diverse scenes, activities, and camera motions.

InFlux++ Synth (Sec. 4.1) builds on Infinigen [19,20] (Sec. 4.1.1) to generate diverse indoor and nature scenes with accurate per-frame annotations, adding camera motion (Sec. 4.1.2), dynamic intrinsics (Sec. 4.1.3, Sec. 4.1.4), and lens distortion with other augmentations (Sec. 4.1.5) to produce a diverse (Sec. 4.1.6) dataset suitable for training dynamic intrinsics prediction models.

InFlux++ Real (Sec. 4.2) extends InFlux with 514K+ frames across 334 high-resolution videos. Ground truth intrinsics are obtained via InFlux’s lens calibration methodology applied to our lenses, with modifications to the procedure for large field of view (FOV) spatial footprint calibration experiments (Sec. 4.2.1). After lens calibration, we collect videos and blur frames containing faces or license plates (Sec. 4.2.2) for privacy. Our benchmark videos are diverse (Sec. 4.2.3), covering a wider range of scenes and camera motions than InFlux.

4.1 InFlux++ Synth

4.1.1 Procedural Scene Generation and Renderer. We build InFlux++ Synth on top of Infinigen [19,20], a procedural scene generator and renderer, for two main reasons. First, Infinigen allows the creation of a virtually limitless variety of indoor and nature scenes, with diverse lighting, materials, and scene composition. Its layout optimizer and tuned parameter distributions ensure scene

realism without any manual design, thus providing scene diversity critical for constructing a training dataset that generalizes well across environments.

Second, because Infinigen renders synthetic videos via Blender, we have full access to underlying camera and scene information, including the rendering camera’s intrinsics. This allows us to generate diverse video sequences with perfect per-frame ground truth intrinsics even in the presence of dynamic objects, providing a level of accurate supervision that would be infeasible with real-world captures. We also provide per-frame camera pose, depth, and surface normals for a subset of the dataset; see our supplement for additional details. These capabilities make InFlux++ Synth well-suited for training models to predict dynamic intrinsics in varied and realistic scenarios.

4.1.2 Camera Motion. To obtain varied camera motion, we build on the Rapidly-Exploring Random Trees (RRT) camera trajectory generator provided by Infinigen [19]. After coarse scene geometry is generated, the planner samples collision-free positional waypoints, assigns camera orientations, and connects the resulting pose keyframes using Bézier interpolation. This results in natural camera motion with smooth translational and rotational motion.

To favor visually informative views with richer geometric structure, we add a surface normal and depth-based criterion that rejects camera orientations that produce viewpoints dominated by a single coherent surface, such as a blank wall or featureless terrain. We also strengthen orientation resampling and introduce local segment replanning, allowing RRT to recover from locally invalid viewpoints without discarding all previous motion. See our supplement for details.

4.1.3 Per-Frame Intrinsics Parameterization. To produce realistic variation of camera intrinsics in our synthetic videos, we parameterize the CFL of the rendering camera using LFL and LTO, rather than directly setting CFL as Blender allows. This approach is inspired by real-world zoom lenses, where the camera intrinsics are fully parameterized by LFL and FD, which are physically adjusted via the zoom and focus rings, respectively.

To render a video with varying intrinsics, we vary the LFL and LTO values of each frame. For now, we focus on computing the CFL for a single frame. Given the LFL and LTO for a frame, we use the thin lens model to compute CFL as a function of these two parameters. This approach is consistent with Blender’s Cycles rendering engine, which performs ray tracing based on the thin lens equation. Specifically, we have that:

$$\begin{aligned}\frac{1}{\text{LFL}} &= \frac{1}{\text{CFL}} + \frac{1}{\text{LTO}} \\ \text{CFL} &= \frac{\text{LFL} \cdot \text{LTO}}{\text{LTO} - \text{LFL}}\end{aligned}$$

By parameterizing CFL in this manner, each rendered frame’s effective focal length changes in a physically meaningful manner. In addition, this formulation naturally produces lens breathing, a visual phenomenon observed in real-world

Algorithm 1 Bounded Random Walk for LFL over Video Frames

```

1: Input:
    $x_{\min}, x_{\max}$ : LFL bounds
    $\delta_{\max}$ : maximum magnitude of LFL change between keyframes
    $s_{\min}, s_{\max}$ : min/max frames between keyframes
    $T$ : total number of video frames
2: Sample initial LFL  $x_0 \sim \text{Uniform}(x_{\min}, x_{\max})$ 
3: Set  $t \leftarrow 0$ ,  $x \leftarrow x_0$  and record first keyframe  $(t, x)$ 
4: while  $t < T$  do
5:   repeat
6:     Sample  $\Delta x \sim \text{Uniform}(-\delta_{\max}, \delta_{\max})$ 
7:      $x_{\text{new}} \leftarrow x + \Delta x$ 
8:   until  $x_{\text{new}} \in [x_{\min}, x_{\max}]$ 
9:   Sample  $\Delta t \sim \text{Uniform}(s_{\min}, s_{\max})$ 
10:   $t_{\text{new}} \leftarrow t + \Delta t$ 
11:  if  $t_{\text{new}} \geq T$  then
12:     $t_{\text{new}} \leftarrow T$  ▷ ensure last frame is a keyframe
13:  end if
14:  Record keyframe:  $(t_{\text{new}}, x_{\text{new}})$ 
15:   $t \leftarrow t_{\text{new}}$ ,  $x \leftarrow x_{\text{new}}$ 
16: end while
17: Apply Bézier interpolation between keyframes to compute LFL for all other frames

```

cameras where the FOV can shift slightly when the LTO changes, even with a fixed LFL. This realistic visual effect does not occur by default in Blender, where CFL and LTO are independently set. By adding lens breathing to our video renders, we provide a potentially valuable supervisory signal for training models to predict dynamic intrinsics. See our supplement for details.

4.1.4 Temporal Variation of LFL and LTO. Given our parameterization of CFL as a function of LFL and LTO, we now describe how we vary LFL and LTO over time to produce realistic changing intrinsics for our synthetic videos.

To vary LFL, we use a bounded random walk. Suppose the initial LFL value is x_0 at frame $t = 0$. We generate a target LFL for a future frame $t' > 0$ by adding a randomly sampled delta to x_0 , ensuring the result stays within the allowed LFL bounds. The frame t' itself is also chosen randomly within a specified range of steps ahead. This process is repeated sequentially across the video, producing a series of target LFL values at selected frames. We then use Bézier interpolation to smoothly fill in the LFL values for all intermediate frames between these targets, producing realistic smooth changes in LFL over time. For our renders, we bound LFL between 8 mm and 100 mm. See Algorithm 1 for a summary of the bounded random walk method.

We follow a similar strategy to vary LTO. However, the visible scene content changes as the camera moves, so directly performing a bounded random walk with fixed limits on the raw LTO could produce low-quality frames in which the entire scene is out of focus. To address this, we estimate the minimum and

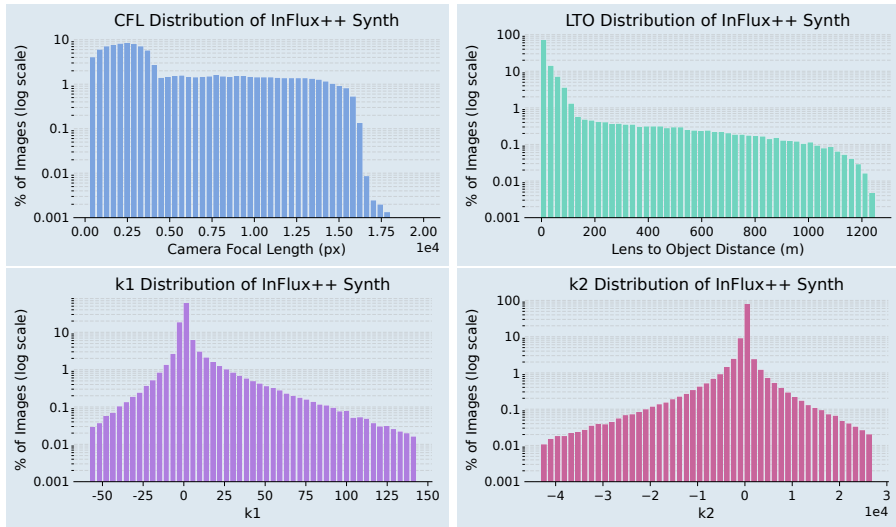


Fig. 4: Histograms of per-frame intrinsics in InFlux++ Synth. Values of k_1 and k_2 are clipped at the 0.1% and 99.9% percentiles for visualization. Both are heavily centered around 0. LTO ranges roughly from 0.04 m to 12 m for indoor scenes and from 0.7 m upward for nature scenes. These value ranges arise naturally from scene geometry.

maximum visible scene depths at each keyframe via ray tracing, denoted as d_{near} and d_{far} , and parameterize the LTO as a fractional position α within this interval. The resulting LTO is computed as $d_{\text{focus}} = d_{\text{near}} + (d_{\text{far}} - d_{\text{near}})\alpha$, where α is constrained between 0 and 1 and changes via a bounded random walk. This formulation produces smooth temporal variation in LTO while ensuring that some portion of the scene remains in focus, even as camera motion alters the visible depth distribution. See our supplement for additional details.

4.1.5 Lens Distortion and Augmentations. The videos produced by Infinigen are by default undistorted. To introduce lens distortion and other data augmentations useful for training dynamic intrinsics prediction models, we provide a data loader based on [29,30] that applies these effects on-the-fly.

For lens distortion, we adopt the Brown–Conrady [2] model. To ensure realistic distortion, we first sample reasonable magnitudes of radial distortion at two control points on the edge of the image and then solve for k_1, k_2 values that produce these effects. We optionally sample small tangential p_1, p_2 values. Our data loader can be further customized; see our supplement for details.

In addition to distortion, we also apply color transforms, resizing transforms, and noising operations following [29]. However, we disable blurring, sharpening, and resolution changes. See our supplement for details.

4.1.6 Data Diversity. We generate 441K+ images across 1841 videos, which are rendered at a resolution of 1280×720 and span 10 seconds at 24 FPS. Of these videos, 1040 are indoor scenes, and 801 are nature scenes.

Camera Intrinsic Diversity. The bounded random walks used to vary LFL and LTO produce smooth, continuous variation in camera intrinsics throughout each sequence. As a result, downstream intrinsics parameterized by LFL and LTO, such as CFL, k_1 , and k_2 , exhibit broad coverage across their feasible ranges rather than clustering around a small set of discrete values. See Fig. 4 for the empirical distributions of these parameters.

Scene Diversity. Our dataset includes a wide range of indoor environments, including living rooms, bedrooms, bathrooms, kitchens, office spaces, and more. These scenes exhibit realistic variation in layout, geometry, lighting conditions, and material properties. In particular, we include glossy and reflective surfaces as well as mirrors to capture challenging real-world optical effects.

Our nature scenes span diverse environments such as forests, deserts, caves, mountains, canyons, coasts, snowy arctic landscapes, coral reefs, and more. These environments are varied in scene layout, scale, geometry, texture, and visibility. Some scenes include dynamic objects, with animals such as birds, dragonflies, insects, snakes, and other terrestrial animals following collision-aware trajectories. Other scenes exhibit natural motion from water waves, ripples, and falling leaves. A qualitative gallery illustrating scene diversity is provided in Fig. 2.

4.2 InFlux++ Real

To construct InFlux++ Real, we collect real-world videos with dynamic intrinsics and per-frame ground truth intrinsics annotations largely following the data collection procedure introduced in InFlux [12]. Like InFlux, we utilize specialized zoom lenses that record lens metadata, including per-frame LFL and FD. Since these values uniquely determine the optical state of the lens, a one-time calibration process can be used to construct a lens-specific lookup table (LUT) mapping (LFL, FD) pairs to camera intrinsics. This allows per-frame intrinsics to be recovered during video capture without interrupting natural camera motion.

To construct a LUT for each lens, we perform camera calibration experiments over a wide grid of LFL and FD values. InFlux divides these experiments into two categories based on the FOV spatial footprint (FSF) of the camera, defined as the 3D region within the camera’s FOV that is in focus. For small to medium FSF cases, we follow InFlux and perform calibration by moving boards of varying sizes throughout the camera’s FSF. For large FSF cases, InFlux employs drone-based calibration, since moving a calibration board through a much larger FSF can be prohibitively difficult. In contrast, we introduce a new board-based calibration procedure that remains effective even for large FSF. This eliminates the need for drone-based experiments and significantly simplifies the data collection process.

Table 1: Scene category video counts for InFlux and InFlux++ Real.

Scene Category	InFlux	InFlux++ Real
INDOOR	113	123
Controlled & Staged	41	14
Workspace & Academic	60	53
Everyday & Domestic	0	22
Casual Activity & Play	6	4
Tabletop & Close-up	6	24
Elevated & Aerial Vantage	0	6
OUTDOOR	273	211
Campus & Architecture	95	20
Nature & Parks	53	3
Urban	0	20
Suburban & Street	91	64
Monuments & Landmarks	9	0
Nighttime	25	23
Sports & Recreation	0	58
Elevated & Aerial Vantage	0	3
Moving Vehicle	0	20
Total	386	334

4.2.1 Large FSF Board-based Calibration. Obtaining high-quality board-based calibration data requires moving the calibration board throughout the camera’s FOV while ensuring that the pattern remains visible, in focus, and observed under a wide range of rotations. This becomes difficult in practice for large FSF settings, which typically occur when the camera is zoomed out and FD is large. The board must be placed far from the camera to stay in focus; consequently, it must also have a large pattern to remain visible to the camera from that distance. However, large boards are cumbersome to move, making it difficult to sufficiently excite rotational axes and cover the full FSF.

Rather than moving a large board around, we instead move our camera around a static large board. This still allows the board to assume a variety of poses and angles relative to the camera—note that there is no requirement for the camera to stay static for calibration. Hence, by moving and rotating the camera, we can cover the full FSF and excite all axes of rotation with the board, which is much easier than moving the board for large FSF settings.

For our large FSF experiments, the calibration target we use is an array of AprilTags [16] projected onto a 5.45 m $\times$ 3.06 m rigid screen. The screen is installed in a multi-story lecture hall with a balcony. We record footage from four camera positions: left and right on the ground floor, and left and right on the balcony. At each position, we vary the camera rotation, effectively moving the board through the camera FOV. This simultaneously covers the full FSF while exciting rotational axes. See our supplement for additional details on the lenses and calibration experiments used to construct the LUTs.

4.2.2 Face and License Plate Blurring. To protect the privacy of subjects in our real-world benchmark, we apply modified versions of [4] and [18] to every frame to detect and blur faces and license plates. See our supplement for details.

4.2.3 Data Diversity. We capture 514K+ images across 334 videos. The resolution of our footage is 3424 $\times$ 2202, filmed at 23.976 FPS.

Scene Diversity. Compared to InFlux, InFlux++ Real exhibits greater diversity in both scene types and activities. InFlux outdoor footage is largely shot in campus environments, suburban streets, and natural settings. In contrast, InFlux++ Real includes a broader range of urban environments and scenes involving active human motion, such as doing chores and sports. It also introduces new types of captures such as moving-vehicle footage and elevated viewpoints. See Tab. 1 for a comparison of video category statistics and Fig. 3 for a gallery.

Camera Motion Diversity. In InFlux, camera motion is mostly rotational. In contrast, InFlux++ Real contains substantially more translational motion, with many sequences captured while the camera is handheld, mounted on a wearable rig, or carried on a moving vehicle. This produces more parallax and viewpoint changes; see our supplement for a qualitative comparison of parallax in InFlux++ Real and InFlux. Because InFlux++ Real videos have extended movement through scenes, its average video length is also longer.

Lens and Intrinsics Diversity. We use the same lens types as [12], but different physical lenses of the same type can have different intrinsics LUTs. One of our lenses has a large principal point offset, expanding the range of c_x and c_y values in InFlux++ Real compared to InFlux. See our supplement for details.

5 Experiments

5.1 Evaluation Metrics and Procedure

To evaluate predicted intrinsics against ground truth, we report recall at several thresholds for CFL and principal point percent error. We also report EPE recall at several thresholds, but modify its computation [12] to ensure metric reliability.

First, to ensure EPE is computed only over physically visible points, we revise EPE’s 3D point visibility filter. The original protocol [12] keeps any candidate 3D point whose 2D projection via ground truth intrinsics and distortion falls within the image. But under extreme Brown-Conrady [2] distortion, the polynomial radial model can become non-monotonic and incorrectly map far-off-axis points back into the image, even though a real lens would not observe these points. We fix this by retaining only points on the main monotonic branch of the radial mapping, cutting off the mapping at a radius r_* . See our supplement for details.

Second, to avoid evaluating EPE recall on InFlux++ Real or InFlux frames whose ground truth intrinsics may be unreliable due to LUT interpolation, we define LUT-reliable EPE recall. For each LUT experiment, we estimate LUT interpolation reliability by performing leave-one-out (LOO) validation [12] and comparing the interpolated and measured intrinsics via EPE recall@ T pixels (px). When computing LUT-reliable EPE recall@ T px, we include only frames whose enclosing LUT region has all vertices with LOO EPE recall@ T px ≥ 0.95 . Unless otherwise specified, all EPE metrics reported in this paper use the LUT-reliable formulation. See our supplement for details.

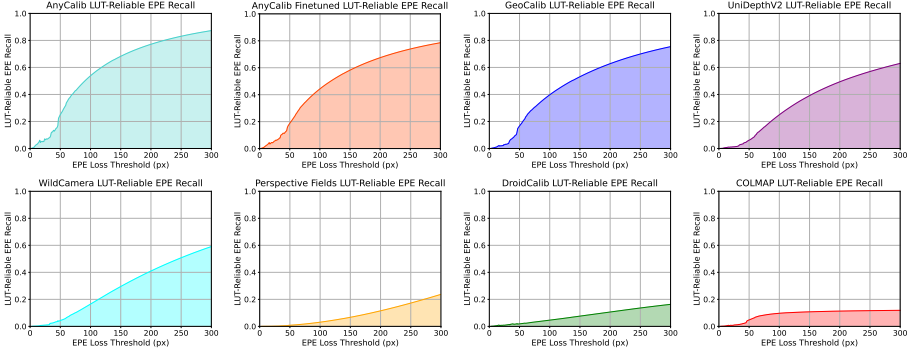


Fig. 5: LUT-reliable EPE recall plots for baseline methods, evaluated on the test splits of InFlux++ Real and InFlux for every integer threshold between 1 px and 300 px. AnyCalib performs the best, but still has low EPE recall for low pixel thresholds.

Table 2: Baseline intrinsics prediction results on the test splits of InFlux and InFlux++ Real. We report recall at multiple thresholds for f_x , f_y , c_x , and c_y percent error, as well as LUT-reliable EPE recall at multiple pixel thresholds. All recalls are reported as percentages; higher is better. Best is **bolded**, second is underlined.

Method	f_x Recall (%)			f_y Recall (%)			c_x Recall (%)			c_y Recall (%)			LUT-Reliable EPE Recall (%)		
	@1%	@10%	@20%	@1%	@10%	@20%	@0.5%	@1%	@2%	@0.5%	@1%	@2%	@10 px	@50 px	@300 px
AnyCalib [29]	3.00	25.2	41.3	2.96	25.2	41.3	41.6	76.5	98.1	24.1	37.4	46.2	2.70	25.1	87.3
AnyCalib (finetuned)	3.68	31.2	50.4	3.65	31.3	50.5	33.0	67.2	95.4	20.4	33.4	44.1	<u>1.93</u>	<u>18.7</u>	<u>78.6</u>
GeoCalib [30]	2.17	19.7	33.5	2.15	19.6	33.4	49.5	82.3	99.4	38.3	40.9	45.0	8.53e-1	16.7	75.4
UniDepthV2 [17]	1.41	14.0	26.6	1.31	13.1	25.7	21.8	45.4	69.9	3.87	9.14	25.5	5.68e-1	6.14	63.0
WildCamera [32]	1.29	12.9	26.6	1.28	13.0	26.5	4.27	8.91	20.7	5.38	10.8	21.3	1.70e-1	4.26	59.1
Perspective Fields [7]	7.32e-1	7.65	16.7	7.07e-1	7.53	16.5	1.48	2.97	5.93	1.55	3.09	6.20	1.29e-1	8.06e-1	23.7
DroidCalib [6]	7.11e-1	5.99	11.8	5.66e-1	5.95	12.0	2.92	6.07	14.4	2.54	4.45	8.48	3.33e-1	1.86	16.3
COLMAP [23]	1.49	6.34	7.22	1.56	6.36	7.23	3.14	7.16	8.87	1.78	1.96	2.31	2.25e-1	4.78	11.9

5.2 Baseline Evaluation using InFlux++ Real and InFlux

We evaluate seven baseline intrinsics prediction methods on the test splits of InFlux++ Real and InFlux: AnyCalib [29], COLMAP [23], DroidCalib [6], GeoCalib [30], Perspective Fields [7], UniDepthV2 [17], and WildCamera [32]. See our supplement for details on the InFlux++ Real test split. For all baselines except AnyCalib (AC), we use the model settings specified in [12]. For AC, we use the `anycalib_gen.pt` weights available on Hugging Face. We report recall@{1, 10, 20}% for f_x and f_y percent errors, recall@{0.5, 1, 2}% for c_x and c_y percent errors, and recall@{10, 50, 300} px for LUT-reliable EPE. See Tab. 2 for a table of baseline performance and Fig. 5 for a plot of LUT-reliable EPE recall at additional thresholds. We additionally provide a breakdown of AC performance on InFlux++ Real versus InFlux in Tab. 3.

Overall, the results show that dynamic intrinsics prediction remains challenging. AC performs best, yet only achieves 25.2% for f_x recall@10% and 25.1% for EPE recall@50 px. AC achieves higher f_x , f_y , and EPE recall on InFlux++ Real than on InFlux, but worse c_x and c_y recall. This may be because InFlux++ Real has proportionally more indoor scenes than InFlux, which better matches AC’s

Table 3: Evaluation of AnyCalib before and after finetuning on InFlux++ Synth. Finetuning improves CFL estimation. Best is **bolded** within each evaluation set.

Eval. Set	Method	f_x Recall (%)			f_y Recall (%)			c_x Recall (%)			c_y Recall (%)			LUT-Reliable EPE Recall (%)		
		@1% @10% @20%	@1% @10% @20%	@1% @10% @20%	@1% @10% @20%	@1% @10% @20%	@1% @10% @20%	@0.5% @1% @2%	@0.5% @1% @2%	@0.5% @1% @2%	@0.5% @1% @2%	@0.5% @1% @2%	@0.5% @1% @2%	@10 px @50 px @300 px	@10 px @50 px @300 px	@10 px @50 px @300 px
InFlux	AnyCalib [29]	1.30	11.5	20.8	1.38	11.7	21.0	73.5	94.4	99.6	51.3	81.7	99.2	2.95	34.1	80.1
	AnyCalib (finetuned)	2.21	17.5	34.1	2.23	18.7	34.9	65.5	90.0	98.5	47.4	76.1	98.3	2.09	27.9	82.7
InFlux++ Real	AnyCalib [29]	3.36	28.2	45.7	3.31	28.1	45.7	34.7	72.6	97.8	18.2	27.8	34.7	6.40e-3	24.3	87.9
	AnyCalib (finetuned)	4.00	34.2	54.0	3.96	34.0	53.9	25.9	62.3	94.7	14.5	24.2	32.3	1.66e-1	17.9	78.2
Combined	AnyCalib [29]	3.00	25.2	41.3	2.96	25.2	41.3	41.6	76.5	98.1	24.1	37.4	46.2	2.70	25.1	87.3
	AnyCalib (finetuned)	3.68	31.2	50.4	3.65	31.3	50.5	33.0	67.2	95.4	20.4	33.4	44.1	1.93	18.7	78.6

indoor-heavy training distribution. InFlux++ Real contains more off-center principal point frames, which could make accurate c_x and c_y prediction harder.

5.3 AnyCalib Finetuning using InFlux++ Synth

We explore the idea of using synthetic supervision for dynamic intrinsics prediction by finetuning AC with a subset of InFlux++ Synth. We perform finetuning for 15 epochs and take the checkpoint with the lowest CFL percent error on the validation split of InFlux++ Real. See our supplement for details.

As shown in Tab. 3, finetuning AC on InFlux++ Synth improves recall for f_x and f_y error at all thresholds. However, it worsens recall for c_x , c_y , and EPE, with the exception of EPE recall@10 px for InFlux++ Real and EPE recall@300 px for InFlux. One possible explanation is that AC’s FOV-field training loss is not directly aligned with EPE, especially near image boundaries where lens distortion is most pronounced. This may limit the model’s ability to fit accurate distortion parameters, which strongly affect EPE; see our supplement for details. Nevertheless, these results suggest that InFlux++ Synth is a promising source of supervision for dynamic intrinsics prediction models.

6 Conclusion

We present InFlux++, a data suite for dynamic intrinsics prediction. InFlux++ Synth provides synthetic videos with accurate and dynamic intrinsics, diverse scenes and materials, dynamic objects, and realistic optical effects such as defocus blur. InFlux++ Real expands upon the existing real-world benchmark InFlux with greater scene and camera motion diversity. Current methods struggle to predict accurate intrinsics on InFlux++ Real and InFlux. However, finetuning with InFlux++ Synth data is a promising direction for improving intrinsics prediction. Through InFlux++, we aim to support and encourage future progress in dynamic intrinsics prediction.

Acknowledgements

Erich Liang was supported in part by the NSF Graduate Research Fellowship Program under Grant No. 2146752. We thank our friends and colleagues at Princeton University for their help with filming the benchmark.

References

- Atienza, R., Zelinsky, A.: A practical zoom camera calibration technique: An application of active vision for human-robot interaction (10 2002)
- Brown, D.C.: Decentering Distortion of Lenses. *Photogrammetric Engineering* **32**(3), 444–462 (1966)
- Burri, M., Nikolic, J., Gohl, P., Schneider, T., Rehder, J., Omari, S., Achtelik, M.W., Siegwart, R.: The euroc micro aerial vehicle datasets. *Int. J. Rob. Res.* **35**(10), 1157–1163 (Sep 2016). <https://doi.org/10.1177/0278364915620033>, <https://doi.org/10.1177/0278364915620033>
- Deng, J., Guo, J., Ververas, E., Kotsia, I., Zafeiriou, S.: Retinaface: Single-shot multi-level face localisation in the wild. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2020)
- Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. *Int. J. Rob. Res.* **32**(11), 1231–1237 (Sep 2013). <https://doi.org/10.1177/0278364913491297>, <https://doi.org/10.1177/0278364913491297>
- Hagemann, A., Knorr, M., Stiller, C.: Deep geometry-aware camera self-calibration from video. In: *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. pp. 3415–3425 (2023). <https://doi.org/10.1109/ICCV51070.2023.00318>
- Jin, L., Zhang, J., Hold-Geoffroy, Y., Wang, O., Blackburn-Matzen, K., Sticha, M., Fouhey, D.F.: Perspective Fields for Single Image Camera Calibration . In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 17307–17316. IEEE Computer Society, Los Alamitos, CA, USA (Jun 2023). <https://doi.org/10.1109/CVPR52729.2023.01660>, <https://doi.ieeecomputersociety.org/10.1109/CVPR52729.2023.01660>
- Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics* **42**(4) (July 2023), <https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/>
- Krishnan, A., Liu, S., Sarlin, P.E., Gentilhomme, O., Caruso, D., Monge, M., Newcombe, R., Engel, J., Pollefeys, M.: Benchmarking egocentric visual-inertial slam at city scale. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (2025)
- Li, W., Saeedi, S., McCormac, J., Clark, R., Tzoumanikas, D., Ye, Q., Huang, Y., Tang, R., Leutenegger, S.: Interiornet: Mega-scale multi-sensor photo-realistic indoor scenes dataset. In: *British Machine Vision Conference (BMVC)* (2018)
- Li, Z., Snavely, N.: Megadepth: Learning single-view depth prediction from internet photos. In: *Computer Vision and Pattern Recognition (CVPR)* (2018)
- Liang, E., Bhattacharjee, R., Dey, S., Moschopoulos, R., Wang, C., Liao, M., Tan, G., Wang, A., Kayan, K., Alexandropoulos, S., Deng, J.: Influx: A benchmark for self-calibration of dynamic intrinsics of video cameras. In: Belgrave, D., Zhang, C., Lin, H., Pascanu, R., Koniusz, P., Ghassemi, M., Chen, N. (eds.) *Advances in Neural Information Processing Systems*. vol. 38. Curran Associates, Inc. (2025), https://proceedings.neurips.cc/paper_files/paper/2025/file/8a8eca190088852067b4e8cc1b907122-Paper-Datasets_and_Benchmarks_Track.pdf
- McCormac, J., Handa, A., Leutenegger, S., Davison, A.J.: Scenenet rgb-d: Can 5m synthetic images beat generic imagenet pre-training on indoor segmentation? In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (Oct 2017)

14. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I. p. 405–421. Springer-Verlag, Berlin, Heidelberg (2020). https://doi.org/10.1007/978-3-030-58452-8_24, https://doi.org/10.1007/978-3-030-58452-8_24
15. Mur-Artal, R., Tardós, J.D.: ORB-SLAM2: an open-source SLAM system for monocular, stereo and RGB-D cameras. *IEEE Transactions on Robotics* **33**(5), 1255–1262 (2017). <https://doi.org/10.1109/TR0.2017.2705103>
16. Olson, E.: AprilTag: A robust and flexible visual fiducial system. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). pp. 3400–3407. IEEE (May 2011)
17. Piccinelli, L., Sakaridis, C., Yang, Y.H., Segu, M., Li, S., Abbeloos, W., Van Gool, L.: Unidepthv2: Universal monocular metric depth estimation made simpler. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **48**(3), 2354–2367 (2026). <https://doi.org/10.1109/TPAMI.2025.3628473>
18. Raina, N., Somasundaram, G., Zheng, K., Miglani, S., Saarinen, S., Meissner, J., Schwesinger, M., Pesqueira, L., Prasad, I., Miller, E., Gupta, P., Yan, M., Newcombe, R., Ren, C., Parkhi, O.M.: Egoblur: Responsible innovation in aria (2023)
19. Raistrick, A., Lipson, L., Ma, Z., Mei, L., Wang, M., Zuo, Y., Kayan, K., Wen, H., Han, B., Wang, Y., Newell, A., Law, H., Goyal, A., Yang, K., Deng, J.: Infinite photorealistic worlds using procedural generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12630–12641 (2023)
20. Raistrick, A., Mei, L., Kayan, K., Yan, D., Zuo, Y., Han, B., Wen, H., Parakh, M., Alexandropoulos, S., Lipson, L., Ma, Z., Deng, J.: Infinigen indoors: Photorealistic indoor scenes using procedural generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 21783–21794 (June 2024)
21. Roberts, G.: Lenses. *Mastering Photography* pp. 41–63 (1995)
22. Sarlin, P.E., Dusmanu, M., Schönberger, J.L., Speciale, P., Gruber, L., Larsson, V., Miksik, O., Pollefeys, M.: LaMAR: Benchmarking Localization and Mapping for Augmented Reality. In: ECCV (2022)
23. Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)
24. Schops, T., Schonberger, J.L., Galliani, S., Sattler, T., Schindler, K., Pollefeys, M., Geiger, A.: A multi-view stereo benchmark with high-resolution images and multi-camera videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
25. Sturm, J., Engelhard, N., Endres, F., Burgard, W., Cremers, D.: A benchmark for the evaluation of rgb-d slam systems. In: 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 573–580 (2012). <https://doi.org/10.1109/IROS.2012.6385773>
26. Teed, Z., Deng, J.: Droid-slam: deep visual slam for monocular, stereo, and rgb-d cameras. In: Proceedings of the 35th International Conference on Neural Information Processing Systems. NIPS '21, Curran Associates Inc., Red Hook, NY, USA (2021)
27. Teed, Z., Lipson, L., Deng, J.: Deep patch visual odometry. *Advances in Neural Information Processing Systems* **36**, 39033–39051 (2023), https://proceedings.neurips.cc/paper_files/paper/2023/file/7ac484b0f1a1719ad5be9aa8c8455fbb-Paper-Conference.pdf

28. Tesch, J., Becherini, G., Achar, P., Yiannakidis, A., Kocabas, M., Patel, P., Black, M.J.: BEDLAM2.0: Synthetic humans and cameras in motion. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2025)
29. Tirado-Garín, J., Civera, J.: Anycalib: On-manifold learning for model-agnostic single-view camera calibration. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 8044–8055 (October 2025)
30. Veicht, A., Sarlin, P.E., Lindenberger, P., Pollefeys, M.: Geocalib: Learning single-image calibration with geometric optimization. In: Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part XL. p. 1–20. Springer-Verlag, Berlin, Heidelberg (2024). https://doi.org/10.1007/978-3-031-73661-2_1, https://doi.org/10.1007/978-3-031-73661-2_1
31. Wang, W., Zhu, D., Wang, X., Hu, Y., Qiu, Y., Wang, C., Hu, Y., Kapoor, A., Scherer, S.: Tartanair: A dataset to push the limits of visual slam. In: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). p. 4909–4916. IEEE Press (2020). <https://doi.org/10.1109/IROS45743.2020.9341801>, <https://doi.org/10.1109/IROS45743.2020.9341801>
32. Zhu, S., Kumar, A., Hu, M., Liu, X.: Tame a wild camera: in-the-wild monocular camera calibration. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. NIPS ’23, Curran Associates Inc., Red Hook, NY, USA (2023)