Shakes on a Plane: Unsupervised Depth Estimation from Unstabilized Photography

Ilya Chugunov   Yuxuan Zhang   Felix Heide

Princeton University

Abstract

Modern mobile burst photography pipelines capture and merge a short sequence of frames to recover an enhanced image, but often disregard the 3D nature of the scene they capture, treating pixel motion between images as a 2D aggregation problem. We show that in a "long-burst", forty-two 12-megapixel RAW frames captured in a two-second sequence, there is enough parallax information from natural hand tremor alone to recover high-quality scene depth. To this end, we devise a test-time optimization approach that fits a neural RGB-D representation to long-burst data and simultaneously estimates scene depth and camera motion. Our plane plus depth model is trained end-to-end, and performs coarse-to-fine refinement by controlling which multi-resolution volume features the network has access to at what time during training. We validate the method experimentally, and demonstrate geometrically accurate depth reconstructions with no additional hardware or separate data pre-processing and pose-estimation steps.

1. Introduction

Over the last century we saw not only the rise and fall in popularity of film and DSLR photography, but of standalone cameras themselves. We've moved into an era of ubiquitous multi-sensor, multi-core, multi-use, mobile-imaging platforms [12]. Modern cellphones offer double-digit megapixel image streams at high framerates; optical image stabilization; on-board motion measurement devices such as accelerometers, gyroscopes, and magnetometers; and, most recently, integrated active depth sensors [43]. This latest addition speaks to a parallel boom in the field of depth imaging and 3D reconstruction [22, 84]. As users often photograph people, plants, food items, and other complex 3D shapes, depth can play a key role in object understanding tasks such as detection, segmentation, and tracking [32, 63, 80]. 3D information can also help compensate for non-ideal camera hardware and imaging settings through scene relighting [20, 55, 79], simulated depth-of-field effects [1, 71, 72], and frame interpolation [2]. Beyond helping improve or understand RGB content, depth itself is a valuable output for simulating objects in augmented reality [5, 13, 44, 64] and interactive experiences [26, 36].

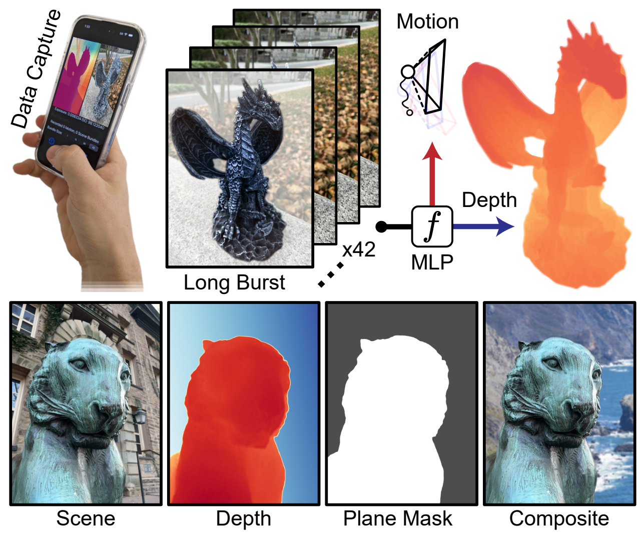


Figure 1. Our neural RGB-D model fits to a single *long-burst* image stack to distill high quality depth and camera motion. The model's *depth-on-a-plane* decomposition can facilitate easy background masking, segmentation, and image compositing.

Depth reconstruction can be broadly divided into *passive* and *active* approaches. *Passive* monocular depth estimation methods leverage training data to learn shape priors [6, 30, 59] – e.g., what image features imply curved versus flat objects or occluding versus occluded structures – but have a hard time generalizing to out-of-distribution scenes [48, 60]. Multi-view depth estimation methods lower this dependence on learned priors by leveraging parallax information from camera motion [16, 69] or multiple cameras [45, 67] to recover geometrically-guided depth. The recent explosion in neural radiance field approaches [49, 50, 66, 81] can be seen a branch of multi-view stereo where a system of explicit geometric constraints is swapped for a more general learned scene model. Rather than classic feature extraction and matching, these models are fit directly to image data to distill dense *implicit* 3D information.

Active depth methods such as pulsed time-of-flight [46] (e.g., LiDAR), correlation time-of-flight [38], and structured light [61, 83] use *controlled illumination* to help with depth reconstruction. While these methods are less reliant on image content than *passive* ones, they also come with complex circuitry and increased power demands [28]. Thus, miniaturization for mobile applications results in very low-resolution *sub-kilopixel sensors* [8, 27, 74]. The Apple iPhone 12-14 Pro devices, which feature one of these miniaturized sensors, use depth derived from RGB, available at 12 *mega-pixel* resolution, to recover scene details lost in the sparse LiDAR measurements. While how exactly they use the RGB stream is unknown, occluding camera sensors reveals that the estimated geometry is the result of *monocular* RGB-guided depth reconstruction.

Returning to the context of mobile imaging, even several seconds of continuous mode photography, which we refer to as a "long-burst", contain only millimeter-scale view variation from natural hand tremor [11]. While these *micro-baseline* [33] shifts are effectively used in burst superresolution and denoising methods [58, 76] as indirect observations of content between sensor pixels, 3D parallax effects on pixel motion are commonly ignored in these models as the depth recovered from this data is too coarse for sub-pixel refinement [31, 33, 82]. A recent work [11] demonstrates high-quality object reconstructions refined with long-burst RGB data, but relies on the iPhone 12 Pro LiDAR sensor for initial depth estimates and device poses, not available on many other cellphones. They treat these poses as ground truth and explicitly solve for depth through minimization of photometric reprojection loss.

In this work, we devise an unsupervised end-to-end approach to jointly estimate high-quality object depth and camera motion from more easily attainable unstabilized two-second captures of 12-megapixel RAW frames and gyroscope data. Our method requires no depth initialization or pose inputs, only a long-burst. We formulate the problem as an image synthesis task, similar to neural radiance methods [50], decomposed into explicit geometric projection through continuous depth and pose models. In contrast to recent neural radiance methods, which typically estimate poses in a pre-processing step, we jointly distill relative depth and pose estimates as a product of simply fitting our model to long-burst data and minimizing photometric loss. In summary, we make the following contributions:

- An end-to-end neural RGB-D scene fitting approach that distills high-fidelity affine depth and camera pose estimates from unstabilized long-burst photography.
- A smartphone data collection application to capture RAW images, camera intrinsics, and gyroscope data for our method, as well as processed RGB frames, low-resolution depth maps, and other camera metadata.
- Evaluations which demonstrate that our approach outperforms existing single and multi-frame image-only depth estimation approaches, with comparisons to high-precision structured light scans to validate the accuracy of our reconstructed object geometries.

Code, data, videos, and additional materials are available on our project website: https://light.princeton.edu/soap

2. Related Work

There exist a wide array of both *active* and *passive* depth estimation methods, ones that recover depth with the help of a controlled illumination source, and ones that use only naturally collected light. We review related work in both categories before discussing neural scene representations.

Active Depth Reconstruction. Structured light and active stereo method rely on patterned illumination to directly infer object shape [15, 83] and/or improve stereo feature matching [61]. In contrast, time-of-flight (ToF) depth sensors use the round trip time of photons themselves – how long it takes light to reach and return from an object – to infer depth. *Indirect* ToF does this by calculating phase changes in continuously modulated light [23, 35, 38], and *direct* ToF times how long a pulse of light is in flight to estimate depth [46,52]. The LiDAR system found in the iPhone 12-14 Pro devices is a type of direct ToF sensor built on low-cost single-photon detectors [8] and solid-state vertical-cavity surface-emitting laser technology [74]. While active LiDAR depth measurements can help produce *metric* depth estimates, without scale ambiguity, existing mobile depth sensors have very limited sub-kilopixel spatial resolution, are sensitive to surface reflectance, and are not commonly found on other mobile devices.

Passive Depth Reconstruction. Single-image passive methods leverage the correlation between visual and geometric features to estimate 3D structure. Examples include depth from shading [3,77], focus cues [78], or generic learned priors [6, 30, 59]. Learned methods have shown great success in producing visually coherent results, but rely heavily on labeled training data and produce unpredictable outputs for out-of-distribution samples. Multi-view and structure from motion works leverage epipolar geometry [24], the relationship between camera and image motion, to extract 3D information from multiple images. Methods typically either directly match RGB features [17, 65], or higher-level learned features [42, 67], in search of depth and/or camera parameters which maximize *photometric consistency* between frames. COLMAP [62] is a widely adopted multi-view method which many neural radiance works [49, 50] rely on for camera pose estimates. In the case of long-burst photography, this problem becomes significantly more challenging as many different depth solutions produce identical images under small view variations. Work in this space either relies on interpolation between sparse feature matches [31, 33, 82] or additional hardware [11] to produce complete depth estimates. Our work builds on these methods to produce both dense depth and high-accuracy camera motion estimates from long-burst image data alone, with a single model trained end-to-end rather than a sequence of disjoint data processing steps.

Neural Scene Representations. Recent work in the area of novel view synthesis has demonstrated that explicit models – e.g. voxel grids, point clouds, or depth maps – are not a necessary backbone to generate high-fidelity representations of 3D space. Rather, the neural radiance family of works, including NeRF [50] and its extensions [4, 10, 53], learn an implicit representation of a 3D scene by fitting a multi-layer perceptron (an MLP) [29] to a set of input images through gradient descent. Similar to multi-view stereo, these methods optimize for photometric loss, ensuring output colors match the underlying RGB data, but they typically don't produce depth maps or camera poses as outputs. On the contrary, most neural radiance methods require camera poses as inputs obtained in a separate pre-processing step from COLMAP [62]. Our setting of long-burst unstabilized photography not only lacks ground truth camera poses, but also provides very little view variation from which to estimate them. While neural scene representation works exist which learn camera poses [41, 73], or operate in the burst photography setting [56], to our best knowledge this is the first work to jointly do both. The most similar recent work by Chugunov et al. [11] uses poses derived from the iPhone 12 Pro ARKit library to learn an implicit representation of depth, but *does not have an image generation model*, and is functionally closer to a direct multi-view stereo approach. In contrast, our work uses a neural representation of RGB as an optimization vehicle to distill high quality continuous representations of both depth and camera poses, with loss backpropogated through an explicit 3D projection model.

3. Long-Burst Photography

Problem Setting. Burst photography refers to the imaging setting where for each button press from the user the camera records multiple frames in rapid succession, sometimes varying parameters such as ISO and exposure time during capture to create a *bracketed sequence* [47]. Burst imaging pipelines investigate how these frames can be merged back into a single higher-fidelity image [12]. These pipelines *typically operate with 2-8 frame captures* and have proven key to high-quality mobile imaging in low-light [25, 40], high dynamic range imaging with low dynamic range sensors [18, 25], and image superresolution, demosaicing, and denoising [75, 76]. On the other end of the imaging spectrum we have video processing literature, which operates on

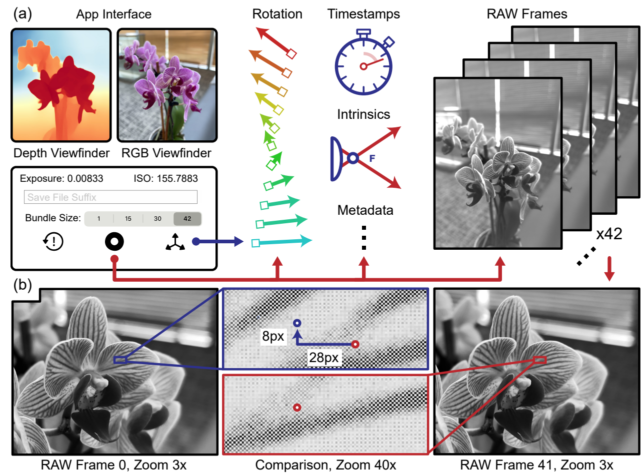


Figure 2. (a) The interface of our app for recording long-burst data. (b) Aligned RAW frames, which illustrate the scale of parallax motion created by natural hand tremor in a two-second capture: a few dozen pixels for an object 30cm from the camera.

sequences hundreds or thousands of frames in length [51] and/or large camera motion [37]. Between these two settings we have what we refer to as "long-burst" photography, several seconds of continuous capture with small view variation. Features built into default mobile camera applications such as Android Motion Photos and Apple Live Photos, which both record three seconds of frames around a button press, demonstrate the ubiquity of *long-burst* data, as they are captured spontaneously without user interaction during natural handheld photography. In this work we capture two-second long-bursts, which result in 42 recorded frames with an average 6mm maximum effective stereo baseline. This produces on the order of several dozen pixels of disparity for close-range objects (<0.5m), see Fig. 2 (b). For an in-depth discussion of motion from natural hand tremor we refer the reader to Chugunov et al. [11].

Data Collection. As there were no commodity mobile applications that allowed for continuous streaming of Bayer RAW frames and metadata, we designed our own data collection tool for long-burst recording. Shown in Fig. 2 (a), it features a live viewfinder with previews of RGB, device depth, and auto-adjusted ISO and exposure values. On a button press, we lock ISO, exposure, and focus, and record a two-second, 42 frame long-burst to the device. Our method uses recorded timestamps, camera intrinsics, gyroscope-driven device rotation estimates, and 12-megapixel RAW frames. However, our app also records processed RGB frames, low-resolution depth maps, and other metadata which we use for validation and visualization.

RAW Images.[1] A modern mobile image signal processing pipeline can have more than a dozen steps between light

[1] *RAW* here refers to sensor data after basic corrections such as compensating for broken and non-uniform pixels, not "raw-raw" data [57].

hitting the CMOS sensor and a photo appearing on screen: denoising, demosaicing, and gamma correction to name a few [12]. While these steps, when finely-tuned, can produce eye-pleasing results, they also pose a problem to downstream computer vision tasks as they break linear noise assumptions, correlate pixel neighborhoods, and lower the overall dynamic range of the content (quantizing the 10- to 14-bit sensor measurements down to 8-bit color depth image files) [7]. In our work we are concerned with the tracking and reconstruction of small image features undergoing small continuous motion from natural hand tremor, and so apply minimal processing to our image data, using linear interpolation to only fill the gaps between Bayer measurements. We preserve the full 14-bit color depth, and fit our depth plus image model directly to this 4032px$\times$3024px$\times$3 channel$\times$42 frame volume.

4. Unsupervised Depth Estimation

In this section, we propose a method for depth estimation from long-burst data. We first lay out the projection model our method relies on, before introducing the scene model, loss functions, and training procedure used to optimize it.

Projection Model. Given an image stack $I(u, v, \text{N})$, where $u, v \in [0, 1]$ are continuous image coordinates and $\text{N} \in [0, 1, ...41]$ is the frame number, we aim to condense the information in $I(u, v, \text{N})$ to a single compact projection model. Given that the motion between frames is small, and image content is largely overlapping, we opt for an RGB-D representation which models each frame of $I(u, v, \text{N})$ as the deformation of some *reference* image $I(u, v)$ projected through depth $D(u, v)$ with a change in camera pose $P(\text{N})$. We expand this process for a single point at coordinates u, v in the reference frame. Let

$$C = [R, G, B]^\top = I(u, v), \quad d = D(u, v) \tag{1}$$

be a sampled colored point C at depth d. Before we can project this point to new frame, we must first convert it from camera (u, v) to world (x, y, z) coordinates. We assume a pinhole camera model to *un-project* this point via

$$\begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} = \boldsymbol{\pi}^{-1}\left(\begin{bmatrix} u \\ v \\ d \end{bmatrix}; K\right) = \begin{bmatrix} d(u - c_x)/f_x \\ d(v - c_y)/f_y \\ d \\ 1 \end{bmatrix}, \tag{2}$$

where K are the corresponding camera intrinsics with focal point (f_x, f_y) and principal point (c_x, c_y). We transform this point from the reference frame to target frame N, with camera pose $P(\text{N})$, via

$$\begin{bmatrix} x^\text{N} \\ y^\text{N} \\ z^\text{N} \end{bmatrix} = [R(\text{N}) \,|\, T(\text{N})] \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} = [P(\text{N})] \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix}. \tag{3}$$

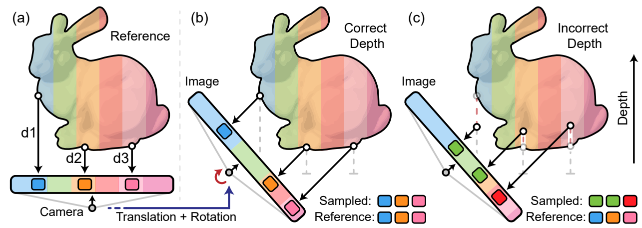


Figure 3. A 2D example of reprojection and sampling. When a reference view (a) is projected to new view with known camera rotation and translation, if the points' depths are accurately estimated they project to sample matching colors in the new image (b). If depths are inaccurate, as in (c), they do not sample corresponding colors, and instead incur *photometric loss*.

Here, $P(\text{N})$ is decomposed into a 3×3 rotation matrix $R(\text{N})$ and 3×1 translation vector $T(\text{N}) = [t_x, t_y, t_z]^\top$. Reverse of the process in Eq. (2), we now *project* this point from the world coordinates $(x^\text{N}, y^\text{N}, z^\text{N})$ in frame N to camera coordinates (u^N, v^N) in the same frame as

$$\begin{bmatrix} u^\text{N} \\ v^\text{N} \end{bmatrix} = \boldsymbol{\pi}\left(\begin{bmatrix} x^\text{N} \\ y^\text{N} \\ z^\text{N} \end{bmatrix}; K(\text{N})\right) = \begin{bmatrix} (f_x^\text{N} x^\text{N})/z^\text{N} + c_x^\text{N} \\ (f_y^\text{N} y^\text{N})/z^\text{N} + c_y^\text{N} \end{bmatrix}, \tag{4}$$

where $K(\text{N})$ are the frame intrinsics. We can now use these coordinates to sample a point from the full image stack

$$C^\text{N} = I(u^\text{N}, v^\text{N}, \text{N}), \quad \mathcal{L}_{photo} = |C - C^\text{N}|. \tag{5}$$

Here $\mathcal{L}_{photo}$ is *photometric loss*, the difference in color between the point we started with in the reference frame and what we sampled from frame N. Given ideal multi-view imaging conditions – no occlusions, imaging noise, or changes in scene lighting – if depth d and pose change $P(\text{N})$ are correct, we will incur no photometric loss $\mathcal{L}_{photo} = 0$ as we sample matching points in both frames. This is visualized in Fig. 3. *Inverting this observation*, we can solve for unknown $D(u, v)$ and $P(\text{N})$ by finding ones that *minimize photometric loss* [62].

Implicit Image Model. In our problem setting, we are given a long-burst image stack $I(u, v, \text{N})$ and device rotation values $R(\text{N})$, supplied by an on-board gyroscope, and are tasked with recovering depth $D(u, v)$ and translation $T(\text{N})$ which make these observations consistent. Given the sheer number of pixels in $I(u, v, \text{N})$, in our case about *500 million*, exhaustively matching and minimizing pixel-to-pixel loss is both computationally intractable and ill-posed. Under small camera motion, many depth solutions for a pixel can map it to identical-colored pixels in the image, especially in textureless parts of the scene. Traditional multi-view stereo (MVS) and bundle adjustment methods tackle this problem with feature extraction and matching [68], optimizing over a *cost-volume* orders of magnitude smaller than the full image space. Here we strongly diverge from

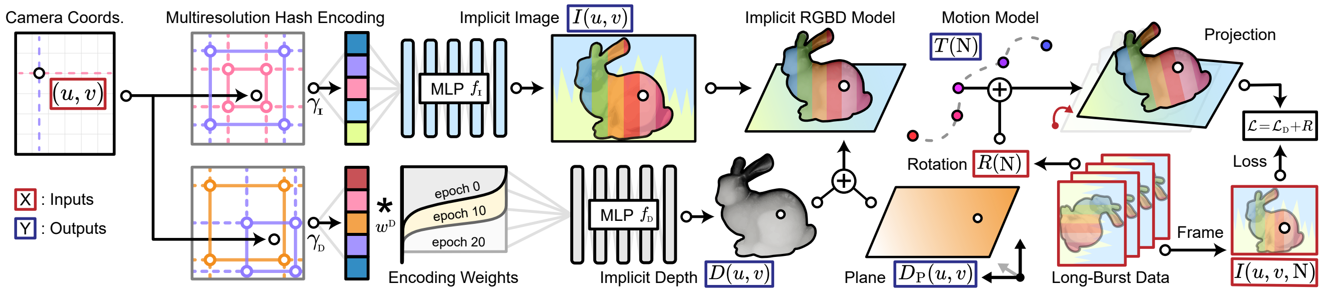


Figure 4. We model a long-burst capture as a single, fully-differentiable forward model comprised of an implicit image $I(u, v)$ projected through implicit depth $D(u, v)$ with motion model $[R(\text{N})|T(\text{N})]$. Calculating reprojection loss with respect to the the captured image stack $I(u, v, \text{N})$, we train this model end-to-end and distill high quality depth and camera motion estimates directly from the burst data.

previous small motion works [11, 31, 82]. Rather than divide the problem into feature extraction and matching, or extract features at all, we propose a single fully differentiable forward model trained *end-to-end*. Depth is distilled as a product of fitting this *neural scene model* to long-burst data. We start by redefining $I(u, v)$ from a static reference image to a learned implicit representation

$$\begin{aligned} I(u, v) &= f_{\text{I}}(\gamma_{\text{I}}(u, v;\ \text{params}_{\gamma\text{I}});\ \theta_{\text{I}}) \\ \text{params}_{\gamma\text{I}} &= \{N_{min}^{\gamma\text{I}},\ N_{max}^{\gamma\text{I}},\ L^{\gamma\text{I}},\ F^{\gamma\text{I}},\ T^{\gamma\text{I}}\} \end{aligned} \tag{6}$$

where f_{I} is a multi-layer perceptron (MLP) [29] with learned weights θ. This MLP learns a mapping from $\gamma_{\text{I}}(u, v)$, a positional encoding of sampled camera coordinates, to image color. Specifically, we borrow the multi-resolution hash encoding from Müller et al. [53] for its spatial aggregation properties. The parameters in $params_{\gamma\text{I}}$ determine the minimum $N_{min}^{\gamma\text{I}}$ and maximum $N_{max}^{\gamma\text{I}}$ grid resolutions, number of grid levels $L^{\gamma\text{I}}$, number of feature dimensions per level $F^{\gamma\text{I}}$, and overall hash table size $T^{\gamma\text{I}}$.

Implicit Depth on a Plane Model. Our depth model is a similar implicit representation with a *learned planar offset*

$$\begin{aligned} d = D(u, v) &= D_{\text{P}}(u, v) + f_{\text{D}}(\gamma_{\text{D}}(u, v; \text{params}_{\gamma\text{D}}); \theta_{\text{D}})^{+} \\ d_{\text{P}} = D_{\text{P}}(u, v) &= \text{a}u + \text{b}v + \text{c}, \end{aligned} \tag{7}$$

where $\{\text{a}, \text{b}, \text{c}\}$ are the learned plane coefficients, and $^{+}$ is the ReLU operation $max(0, x)$. Here $D_{\text{P}}(u, v)$ acts as the depth of the scene background – the surface on or in front of which objects are placed – which is often devoid of parallax cues. Then f_{D} reconstructs the depth of the scene foreground content recovered from parallax in $I(u, v, \text{N})$. While it may seem that we are *increasing* the complexity of the problem, as we now have to learn $I(u, v)$ in addition to $D(u, v)$, this model actually simplifies the learning task when compared to a static $I(u, v)$. Rather than solving for a perfect image from the get-go, f_{I} can move between intermediate representations of the scene with blurry, noisy, and misaligned content, and is gradually refined during training.

Camera Motion Model. Given the continuous, smooth, low-velocity motion observed in natural hand tremor [11], we opt for a low-parameter Bézier curve motion model

$$\begin{aligned} T(\text{N}) = \text{B}(\text{N}; \mathbf{P}^{\text{T}}, N_c^{\text{T}}),\ R(\text{N}) = R_d(\text{N}) + \eta_{\text{R}}\text{B}(\text{N}; \mathbf{P}^{\text{R}}, N_c^{\text{R}}) \\ \text{B}(t; \mathbf{P}, N_c) = \sum_{i=0}^{N_c} \binom{N_c}{i} (1-t)^{N_c - i} t^i \mathbf{P}_i, \end{aligned} \tag{8}$$

with N_c number of control points $\mathbf{P}_i$. Translation estimates $T(\text{N})$ are learned from scratch, whereas rotations $R(\text{N})$ are initialized as device values $R_d(\text{N})$ with learned offsets weighted by η_{R}. Under the small angle approximation [31], we parameterize the rotational offsets $\mathbf{P}^{\text{R}}$ as

$$\mathbf{P}_{\mathbf{i}}^{\text{R}} = \begin{bmatrix} 0 & -r^z & r^y \\ r^z & 0 & -r^x \\ -r^y & r^x & 0 \end{bmatrix}. \tag{9}$$

The choice of N_c controls the dimensionality of the curve on which motion lies – e.g. $N_c = 1$ restricts motion to be linear, $N_c = 2$ is quadratic, and $N_c = 42$ trivially overfits the data with a control point for each frame.

Loss and Regularization. Putting all of the above together we arrive at the full forward model, illustrated in Fig. 4. Given that all of our operations – from re-projection to Bézier interpolation – are fully differentiable, *we train all these components simultaneously, end-to-end, through stochastic gradient descent*. But to do this, we need an objective to minimize. We employ a weighted composite loss

$$\mathcal{L} = \mathcal{L}_{\text{D}} + \alpha_{\text{P}}(\mathcal{L}_{\text{P}}/\mathcal{L}_{\text{D}})\mathcal{R}, \quad \alpha_{\text{P}} > 0,\ \beta_{\text{P}} \geq 1 \tag{10}$$

$$\mathcal{L}_{\text{D}} = |(C - C_{\text{D}}^{\text{N}})/(\text{sg}(C) + \epsilon_{\text{C}})|^2 \tag{11}$$

$$\mathcal{L}_{\text{P}} = |(C - C_{\text{D}}^{\text{P}})/(\text{sg}(C) + \epsilon_{\text{C}})|^2 \tag{12}$$

$$\mathcal{R} = |1 - d/d_{\text{P}}|^2 \tag{13}$$

$$C = I(u, v), \quad C^{\text{N}} = I(u, v, \text{N}), \quad C_{\text{P}}^{\text{N}} = I(u, v, \text{N})_{\text{P}}$$

Here d is the depth output by our combined depth model, and d_{P} is the depth of only the planar component as in Eq. (7). C is a colored point sampled from our implicit image model, C^{N} is the point sampled from the image stack $I(u, v, \text{N})$ following Eqs. (1)–(5) for depth $d = d$, and C_{P}^{N} is the point sampled following Eqs. (1)–(5) for the plane

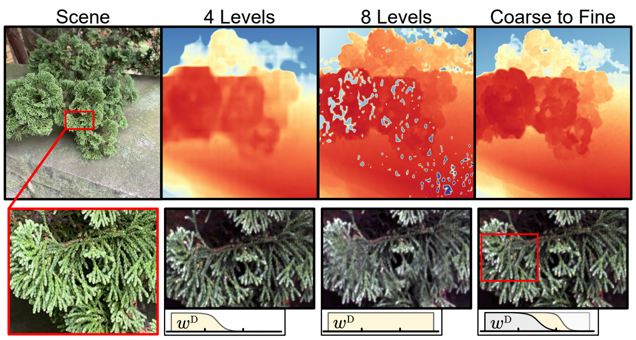


Figure 5. Ablation experiment on the effect of masked encoding levels. Using 4 encoding levels for depth leads to expected lower-resolution depth reconstructions. However, in *8 Levels* where weights are not swept from coarse-to-fine, the reconstruction acquires sharp depth artifacts due to a positive feedback loop during training: high-frequency image gradients from $I(u, v)$ produce discontinuities in estimated depth $D(u, v)$, which in turn produce high-frequency images gradients in $I(u, v)$.

depth $d = d_{\text{P}}$. The regularization term (13) penalizes the magnitude of f_{D}, pulling the depth output towards the plane model. Losses (11) and (12) are relative square photometric errors between sampled colored points, where sg is the stop-gradient operator preventing the denominator C's gradient from being back-propagated. This normalization by the approximate luminance of sampled points is effective in aiding the unbiased reconstruction of HDR images [39], and we refer the reader to derivations in Mildenhall et. al [49] on its relation to tone-mapping. In (10), we combine the photometric loss term $\mathcal{L}_{\text{D}}$, which seeks to maximize overall image reconstruction quality, with a weighted regularization R which penalizes divergence from the planar model. When $\mathcal{L}_{\text{P}} \approx \mathcal{L}_{\text{D}}$ – i.e. the depth offset from f_d is not improving reconstruction quality over a simple plane – the model is strongly penalized. As $\mathcal{L}_{\text{D}}$ decreases relative to $\mathcal{L}_{\text{P}}$ – the implicit depth model f_d improves reconstruction quality – this penalty falls off. In this way, regions that are blurred, textureless, or otherwise have no meaningful parallax information are pulled towards a spatially consistent plane solution rather than spurious depth predictions from f_d. As otherwise, in these regions, there is no photometric penalty for incorrect and noisy depth estimates. The parameter α_{P} controls the strength of this regularization.

Coarse-to-Fine Reconstruction. First estimating low-resolution depths for whole objects before refining features such as edges and internal structures is a tried-and-true technique for improving depth reconstruction quality and consistency [9, 14]. However, one typical caveat of implicit scene representations is the difficulty of performing spatial aggregation – an image pyramid is not well-defined for a continuous representation with no concept of pixel neighborhoods. Rather than try to aggregate outputs, we recognize that the multi-resolution hash encoding $\gamma_{\text{D}}(u, v)$ gives us control over the scale of reconstructed features. By masking the encoding $w^{\text{D}}\gamma_{\text{D}}(u, v)$ with weights $w_i^{\text{D}} \in [0, 1]$ we can restrict the effective spatial resolution of the implicit depth network f_d, as two coordinates that map to the same masked encoding are treated as identical points by f_d. During training, we evolve this weight vector as

$$
\begin{aligned}
w_i^{\text{D}} &= 1/(1 + exp(-ki)) \\
k &= -k_{min} + (\text{epoch} \cdot k_{max})/\text{max\_epochs}
\end{aligned}
\tag{14}
$$

which smoothly sweeps from passing only low-resolution grid encodings to all grid encodings during training, with k_{min} and k_{max} controlling the rate of this sweep. The effects of this masking are visualized in Fig. 5.

Training and Implementation Details. For simplicity of notation we have so far only worked with a single projected point. In practice, during a single forward pass of the model we perform *one-to-all* projection of a batch of 1024 points at a time from the reference $I(u, v)$ to *all 42 frames* in $I(u, v, \text{N})$. We perform stochastic gradient descent on $\mathcal{L}$ with the Adam optimizer [34]. Our implementation is built on tiny-cuda-dnn [54], and on a single Nvidia A100 GPU has a training time of approximately 15 minutes per scene. Our encoding parameters are $N_{min}^{\gamma\text{I}} = 8, N_{max}^{\gamma\text{I}} = 2048, L^{\gamma\text{I}} = 16, F^{\gamma\text{I}} = 4, T^{\gamma\text{I}} = 2^{22}$ and $N_{min}^{\gamma\text{D}} = 8, N_{max}^{\gamma\text{D}} = 128, L^{\gamma\text{D}} = 8, F^{\gamma\text{D}} = 4, T^{\gamma\text{D}} = 2^{14}$, as depth has significantly less high-frequency features than $I(u, v)$. The networks f_{I} and f_{D} are both 5-layer 128 neuron MLPs with ReLU activations. For the rotation offset weight we choose $\eta_{\text{R}} = 10^{-4}$; regularization weight $\alpha_{\text{P}} = 10^{-4}$ and $\epsilon_{\text{C}} = 10^{-3}$; encoding weight parameters $k_{min} = -100, k_{max} = 200$; and number of control points $N_c^{\text{T}} = N_c^{\text{R}} = 21$, one for every two frames. We provide additional training details, and an extensive set of ablation experiments in the Supplemental Document to illustrate the effects of these parameters and how the above values were chosen. Our data capture app is built on the AVFoundation library in iOS 16 and tested with iPhone 12-, 13-, and 14-Pro devices. For consistency, a single 14 Pro device was used for all data captured in this work. RAW capture is hardware/API limited to ~21 FPS, hence a two-second long-burst contains 42 frames.

5. Assessment

Evaluation. We compare our approach to the most similar purely multi-view methods BARF [41] and Depth From Uncalibrated Small Motion Clip (DfUSMC) [21], both of which estimate depth and camera motion directly from an input image stack. We note that BARF also has a similar implicit image generation model. We also compare to learned monocular methods: iPhone's 14 Pro's native depth output and MiDaS [60], a robust single-image approach. Lastly, we compare to Robust Consistent Video Depth Interpolation (RCVD) [37] and Handheld Multi-frame Neural Depth

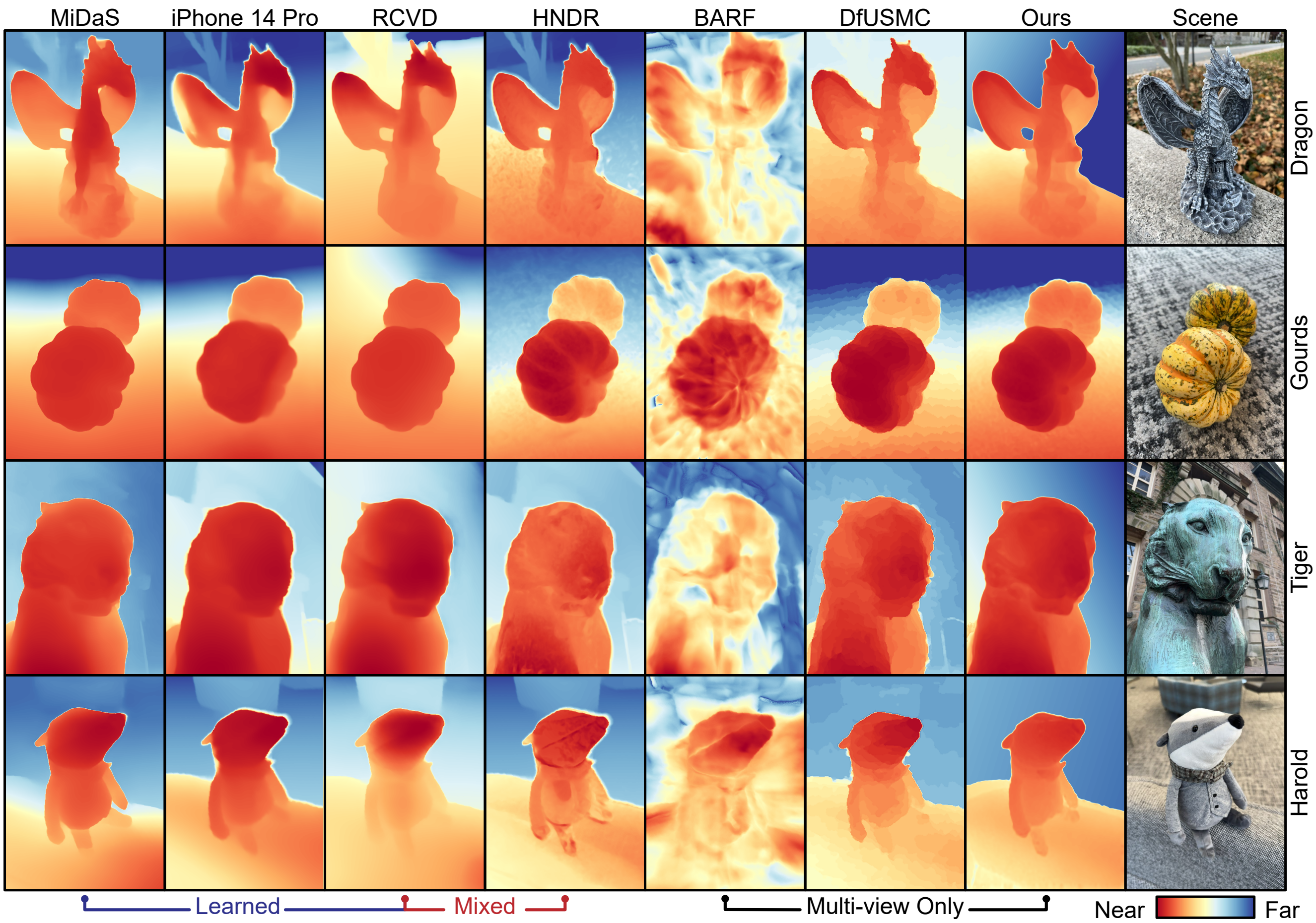
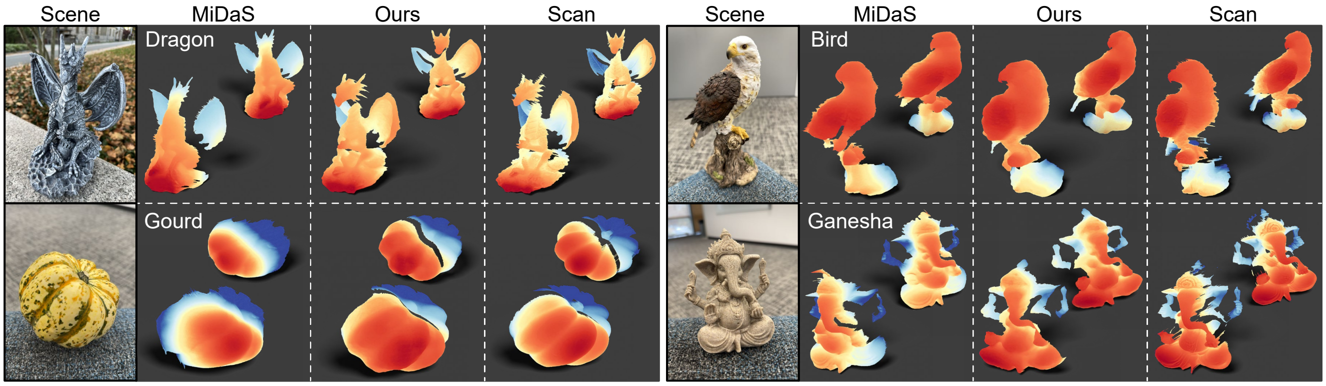


Figure 6. Qualitative comparison of reconstruction results of indoor and outdoor scenes for a range of learned, purely multi-view, and mixed depth estimation methods. Given the mix of depth scales, results are re-scaled by minimizing relative mean square error.

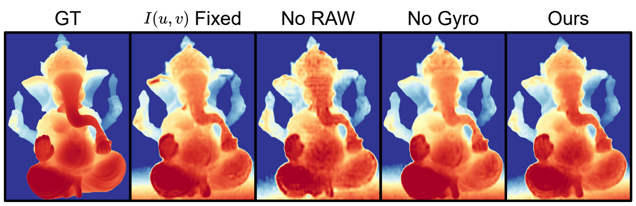
Refinement (HNDR) [11], which both use multi-view information to refine initial depth estimates initialized from a learned monocular approach. The latter of which is most directly related to our approach as it targets close-range objects imaged with multi-view information from natural hand tremor, but relies on iPhone LiDAR hardware for depth initialization and pose estimation. All baselines were run on processed RGB data synchronously acquired by our data capture app, except for HNDR which required its own data capture software that we ran in tandem to ours. We note that other neural scene volume methods such as [49] require COLMAP as a pre-processing step, which *fails to find pose solutions for our long-burst data*. To assess absolute performance and geometric consistency, we scan a select set of complex 3D objects, illustrated in Fig. 7, with a commercial high-precision turntable structured light scanner (Einscan SP). We use this data to generate ground truth object meshes, which we register and render to depth with matching camera parameters to the real captures. For quantitative depth assessment, we use relative absolute error and scale invariant error, commonly used in monocular depth literature [70]; see the Supplemental Document for details.

Reconstruction Quality. Tested on a variety of scenes, illustrated in Fig. 6, we demonstrate high-quality object depth reconstruction outperforming existing learned, mixed, and multi-view only methods. Of particular note is how we are able to reconstruct small features such as *Dragon*'s tail, *Harold*'s scarf, and the ear of the *Tiger* statue consistent to the underlying scene geometry. This is in contrast to methods such as RCVD or HNDR which either neglect to reconstruct the *Tiger*'s ear or reconstruct it behind its head. Our coarse-to-fine approach also allows us to reconstruct scenes with larger low-texture regions, such as *Harold*'s head, which produces striped depth artifacts for HNDR as it can only refine depth within a patch-size of high-contrast edges. Our depth on a plane decomposition avoids spurious depth solutions in low-parallax regions around objects, cleanly segmenting them from their background. This plane segmentation, and it's applications to image and depth matting, are further discussed in the Supplemental Document. In contrast to DfUSMC, which relies on sparse feature matches and RGB-guided filtering to in-paint contiguous depth regions, our unified end-to-end model *directly* produces complete and continuous depth maps.



	MiDaS	iPhone	RCVD	HNDR	BARF	SfSM	Ours		MiDaS	iPhone	RCVD	HNDR	BARF	DfUSMC	Ours
Dragon	.447/.224	.482/.355	.485/.237	.470/.345	.285/.339	.166/.118	**.129**/**.073**	*Bird*	.283/.199	.321/.282	.322/.255	.284/.274	.172/.159	.125/.161	**.082**/**.058**
Gourd	.233/.195	.266/.229	.235/.181	.264/.225	.821/.369	.167/.141	**.086**/**.078**	*Ganesha*	.232/.194	.283/.224	.306/.239	.275/.230	.376/.328	.132/.176	**.094**/**.104**

Figure 7. Object reconstructions visualized as rendered meshes, with associated depth metrics formatted as *relative absolute error / scale invariant error*. Edges over $10\times$ the length of their neighbors were culled to avoid connecting mesh features in occluded regions.



	Fixed	No RAW	No Gyro	Ours		Fixed	No RAW	No Gyro	Ours
Dragon	.150	.142	**.125**	.129	Bird	.113	.148	.089	**.082**
	.081	.097	.080	**.073**		.099	.122	.078	**.058**
Gourd	.091	.093	.088	**.086**	Ganesha	.141	.111	.121	**.094**
	.078	.081	.080	**.078**		.117	.104	**.094**	.104

Figure 8. Ablation study on the effects of fixing the image representation $I(u, v)$ to be the first long-burst frame $I(u, v, 0)$, replacing the RAW data in $I(u, v, \mathrm{N})$ with processed 8-bit RGB, or removing device initial rotation estimates from our model. Metrics formatted as *relative absolute error / scale invariant error*.

In Fig. 7, we highlight our method's ability reconstruct complex objects. While from a *single image* the learned monocular method MiDaS produces visually consistent depth results, from a *single long-burst* our approach directly solves for geometrically accurate affine depth. This difference is most clearly seen in the *Dragon* object, whose wings are reconstructed at completely incorrect depths by MiDAS, disjoint from the rest of the object. This improved object reconstruction is also reflected in the quantitative depth metrics, in which we outperform all comparison methods. Another note is that the most structurally similar method to ours, BARF – which also learns an implicit image model and distills camera poses in a coarse-to-fine encoding approach – fails to produce reasonable reconstructions. We suspect this is related to the findings of Gao et al. [19], that NeRFs do not necessarily obey projective geometry during view synthesis for highly overlapping image content.

Implicit Values of a Learned RAW Model. In Fig. 8, we observe the quantitative and qualitative effects of removing various key method components. For the *No Gyro* tests we replace device rotations $R_d(\mathrm{N})$ with identity rotation for all frames N and learn offsets as usual. We find that while the use of a fixed reference image, 8-bit RGB, or no gyro measurements can reduce our model's average reconstruction quality, all these experiments still converge to acceptable depth solutions. This is especially true of the *No Gyro* experiments, which for many scenes results in *near identical* reconstructions. This further validates our model's ability to independently learn high quality camera pose estimates, and demonstrates its modularity with respect to input data and optimization settings – applicable even to settings where RAW images and device motion data are not available.

6. Discussion and Future Work

In this work, we demonstrate that from only a stack of images acquired during long-burst photography, with parallax information from natural hand-tremor, it is possible to recover high-quality, geometrically-accurate object depth.

Forward Models. Our static single-plane RGB-D representation could potentially be modified to include differentiable models of object motion, deformation, or occlusion.

Image Refinement. We use the learned image $I(u, v)$ as a vehicle for depth optimization, but it could be possible to flip this and use the learned depth $D(u, v)$ as a vehicle for aggregating RGB content (e.g., denoising or deblurring).

From Pixels to Features. Low-texture or distant image regions have insufficient parallax cues for ray-based depth estimation. Learned local feature embeddings could help aggregate spatial information for improved reconstruction.

Acknowledgements. We thank Jinglun Gao and Jun Hu for their support in developing the data capture app. Ilya Chugunov was supported by NSF GRFP (2039656). Felix Heide was supported by a Packard Foundation Fellowship, an NSF CAREER Award (2047359), a Sloan Research Fellowship, a Sony Young Faculty Award, a Project X Innovation Award, and an Amazon Science Research Award.

References

[1] Abdullah Abuolaim and Michael S Brown. Defocus deblurring using dual-pixel data. In *European Conference on Computer Vision*, pages 111–126. Springer, 2020. 1

[2] Wenbo Bao, Wei-Sheng Lai, Chao Ma, Xiaoyun Zhang, Zhiyong Gao, and Ming-Hsuan Yang. Depth-aware video frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3703–3712, 2019. 1

[3] Jonathan T Barron and Jitendra Malik. Shape, illumination, and reflectance from shading. *IEEE transactions on pattern analysis and machine intelligence*, 37(8):1670–1687, 2014. 2

[4] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5855–5864, 2021. 3

[5] Tobias Bertel, Mingze Yuan, Reuben Lindroos, and Christian Richardt. Omniphotos: casual 360 vr photography. *ACM Transactions on Graphics (TOG)*, 39(6):1–12, 2020. 1

[6] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4009–4018, 2021. 1, 2

[7] Tim Brooks, Ben Mildenhall, Tianfan Xue, Jiawen Chen, Dillon Sharlet, and Jonathan T Barron. Unprocessing images for learned raw denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11036–11045, 2019. 4

[8] Clara Callenberg, Zheng Shi, Felix Heide, and Matthias B Hullin. Low-cost spad sensing for non-line-of-sight tracking, material classification and depth imaging. *ACM Transactions on Graphics (TOG)*, 40(4):1–12, 2021. 2

[9] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5410–5418, 2018. 6

[10] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14124–14133, 2021. 3

[11] Ilya Chugunov, Yuxuan Zhang, Zhihao Xia, Xuaner Zhang, Jiawen Chen, and Felix Heide. The implicit values of a good hand shake: Handheld multi-frame neural depth refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2852–2862, 2022. 2, 3, 5, 7

[12] Mauricio Delbracio, Damien Kelly, Michael S Brown, and Peyman Milanfar. Mobile computational photography: A tour. *arXiv preprint arXiv:2102.09000*, 2021. 1, 3, 4

[13] Ruofei Du, Eric Turner, Maksym Dzitsiuk, Luca Prasso, Ivo Duarte, Jason Dourgarian, Joao Afonso, Jose Pascoal, Josh Gladstone, Nuno Cruces, et al. Depthlab: Real-time 3d interaction with depth maps for mobile augmented reality. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, pages 829–843, 2020. 1

[14] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Adv. Neural Inform. Process. Syst.*, pages 2366–2374, 2014. 6

[15] Sean Ryan Fanello, Christoph Rhemann, Vladimir Tankovich, Adarsh Kowdle, Sergio Orts Escolano, David Kim, and Shahram Izadi. Hyperdepth: Learning depth from structured light without matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5441–5450, 2016. 2

[16] Michaël Fonder, Damien Ernst, and Marc Van Droogenbroeck. M4depth: A motion-based approach for monocular depth estimation on video sequences. *arXiv preprint arXiv:2105.09847*, 2021. 1

[17] Silvano Galliani, Katrin Lasinger, and Konrad Schindler. Massively parallel multiview stereopsis by surface normal diffusion. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 873–881, 2015. 2

[18] Orazio Gallo, Alejandro Troccoli, Jun Hu, Kari Pulli, and Jan Kautz. Locally non-rigid registration for mobile hdr photography. In *Proceedings of the IEEE conference on computer vision and pattern recognition Workshops*, pages 49–56, 2015. 3

[19] Hang Gao, Ruilong Li, Shubham Tulsiani, Bryan Russell, and Angjoo Kanazawa. Monocular dynamic view synthesis: A reality check. *arXiv preprint arXiv:2210.13445*, 2022. 8

[20] Kaiwen Guo, Peter Lincoln, Philip Davidson, Jay Busch, Xueming Yu, Matt Whalen, Geoff Harvey, Sergio Orts-Escolano, Rohit Pandey, Jason Dourgarian, et al. The relightables: Volumetric performance capture of humans with realistic relighting. *ACM Transactions on Graphics (ToG)*, 38(6):1–19, 2019. 1

[21] Hyowon Ha, Sunghoon Im, Jaesik Park, Hae-Gon Jeon, and In So Kweon. High-quality depth from uncalibrated small motion clip. In *Proceedings of the IEEE conference on computer vision and pattern Recognition*, pages 5413–5421, 2016. 6

[22] Xian-Feng Han, Hamid Laga, and Mohammed Bennamoun. Image-based 3d object reconstruction: State-of-the-art and trends in the deep learning era. *IEEE transactions on pattern analysis and machine intelligence*, 43(5):1578–1604, 2019. 1

[23] Miles Hansard, Seungkyu Lee, Ouk Choi, and Radu Patrice Horaud. *Time-of-flight cameras: principles, methods and applications*. Springer Science & Business Media, 2012. 2

[24] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 2

[25] Samuel W Hasinoff, Dillon Sharlet, Ryan Geiss, Andrew Adams, Jonathan T Barron, Florian Kainz, Jiawen Chen, and Marc Levoy. Burst photography for high dynamic range and low-light imaging on mobile cameras. *ACM Transactions on Graphics (ToG)*, 35(6):1–12, 2016. 3

[26] Peter Hedman and Johannes Kopf. Instant 3d photography. *ACM Transactions on Graphics (TOG)*, 37(4):1–12, 2018. 1

[27] Eric Hegblom, Yeyu Zhu, Jun Yang, Kelvin Zhang, Benjamin Kesler, Lucas Morales, Matthew Peters, and Jay Skidmore. Column-addressable and matrix-addressable multi-junction vcsel arrays for all electronic-scanning lidar. In *Vertical-Cavity Surface-Emitting Lasers XXVI*, volume 12020, pages 33–41. SPIE, 2022. 2

[28] Radu Horaud, Miles Hansard, Georgios Evangelidis, and Clément Ménier. An overview of depth cameras and range scanners based on time-of-flight technologies. *Machine vision and applications*, 27(7):1005–1020, 2016. 2

[29] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989. 3, 5

[30] Junjie Hu, Mete Ozay, Yan Zhang, and Takayuki Okatani. Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1043–1051. IEEE, 2019. 1, 2

[31] Sunghoon Im, Hyowon Ha, Gyeongmin Choe, Hae-Gon Jeon, Kyungdon Joo, and In So Kweon. High quality structure from small motion for rolling shutter cameras. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 837–845, 2015. 2, 3, 5

[32] Wei Ji, Jingjing Li, Shuang Yu, Miao Zhang, Yongri Piao, Shunyu Yao, Qi Bi, Kai Ma, Yefeng Zheng, Huchuan Lu, et al. Calibrated rgb-d salient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9471–9481, 2021. 1

[33] Neel Joshi and C Lawrence Zitnick. Micro-baseline stereo. *Technical Report MSR-TR-2014–73*, page 8, 2014. 2, 3

[34] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[35] Andreas Kolb, Erhardt Barth, Reinhard Koch, and Rasmus Larsen. Time-of-flight cameras in computer graphics. In *Computer Graphics Forum*, volume 29, pages 141–159. Wiley Online Library, 2010. 2

[36] Johannes Kopf, Kevin Matzen, Suhib Alsisan, Ocean Quigley, Francis Ge, Yangming Chong, Josh Patterson, Jan-Michael Frahm, Shu Wu, Matthew Yu, et al. One shot 3d photography. *ACM Transactions on Graphics (TOG)*, 39(4):76–1, 2020. 1

[37] Johannes Kopf, Xuejian Rong, and Jia-Bin Huang. Robust consistent video depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1611–1621, 2021. 3, 6

[38] Robert Lange. 3d time-of-flight distance measurement with custom solid-state image sensors in cmos/ccd-technology. 2000. 2

[39] Jaakko Lehtinen, Jacob Munkberg, Jon Hasselgren, Samuli Laine, Tero Karras, Miika Aittala, and Timo Aila. Noise2noise: Learning image restoration without clean data. *arXiv preprint arXiv:1803.04189*, 2018. 6

[40] Orly Liba, Kiran Murthy, Yun-Ta Tsai, Tim Brooks, Tianfan Xue, Nikhil Karnad, Qiurui He, Jonathan T Barron, Dillon Sharlet, Ryan Geiss, et al. Handheld mobile photography in very low light. *ACM Trans. Graph.*, 38(6):164–1, 2019. 3

[41] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. Barf: Bundle-adjusting neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5741–5751, 2021. 3, 6

[42] Lahav Lipson, Zachary Teed, and Jia Deng. Raft-stereo: Multilevel recurrent field transforms for stereo matching. *arXiv preprint arXiv:2109.07547*, 2021. 2

[43] Gregor Luetzenburg, Aart Kroon, and Anders A Bjørk. Evaluation of the apple iphone 12 pro lidar for an application in geosciences. *Scientific reports*, 11(1):1–9, 2021. 1

[44] Xuan Luo, Jia-Bin Huang, Richard Szeliski, Kevin Matzen, and Johannes Kopf. Consistent video depth estimation. *ACM Transactions on Graphics (TOG)*, 39(4):71–1, 2020. 1

[45] David Marr and Tomaso Poggio. Cooperative computation of stereo disparity: A cooperative algorithm is derived for extracting disparity information from stereo image pairs. *Science*, 194(4262):283–287, 1976. 1

[46] Aongus McCarthy, Robert J Collins, Nils J Krichel, Verónica Fernández, Andrew M Wallace, and Gerald S Buller. Long-range time-of-flight scanning sensor based on high-speed time-correlated single-photon counting. *Applied optics*, 48(32):6241–6251, 2009. 2

[47] Tom Mertens, Jan Kautz, and Frank Van Reeth. Exposure fusion: A simple and practical alternative to high dynamic range photography. In *Computer graphics forum*, volume 28, pages 161–171. Wiley Online Library, 2009. 3

[48] S Mahdi H Miangoleh, Sebastian Dille, Long Mai, Sylvain Paris, and Yagiz Aksoy. Boosting monocular depth estimation models to high-resolution via content-adaptive multi-resolution merging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9685–9694, 2021. 1

[49] Ben Mildenhall, Peter Hedman, Ricardo Martin-Brualla, Pratul P Srinivasan, and Jonathan T Barron. Nerf in the dark: High dynamic range view synthesis from noisy raw images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16190–16199, 2022. 1, 2, 6, 7

[50] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European conference on computer vision*, pages 405–421. Springer, 2020. 1, 2, 3

[51] Kristina Monakhova, Stephan R Richter, Laura Waller, and Vladlen Koltun. Dancing under the stars: video denoising in starlight. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16241–16251, 2022. 3

[52] Kazuhiro Morimoto, Andrei Ardelean, Ming-Lo Wu, Arin Can Ulku, Ivan Michel Antolovic, Claudio Bruschini, and Edoardo Charbon. Megapixel time-gated spad image sensor for 2d and 3d imaging applications. *Optica*, 7(4):346–354, 2020. 2

[53] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *arXiv preprint arXiv:2201.05989*, 2022. 3, 5

[54] Thomas Müller, Fabrice Rousselle, Jan Novák, and Alexander Keller. Real-time neural radiance caching for path tracing. *arXiv preprint arXiv:2106.12372*, 2021. 6

[55] Rohit Pandey, Sergio Orts Escolano, Chloe Legendre, Christian Haene, Sofien Bouaziz, Christoph Rhemann, Paul Debevec, and Sean Fanello. Total relighting: learning to relight portraits for background replacement. *ACM Transactions on Graphics (TOG)*, 40(4):1–21, 2021. 1

[56] Naama Pearl, Tali Treibitz, and Simon Korman. Nan: Noise-aware nerfs for burst-denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12672–12681, 2022. 3

[57] Kari Pulli. Lecture 11: Camera processing pipeline. *CS231M: Mobile Computer Vision*, May 2015. 3

[58] Guocheng Qian, Yuanhao Wang, Chao Dong, Jimmy S Ren, Wolfgang Heidrich, Bernard Ghanem, and Jinjin Gu. Rethinking the pipeline of demosaicing, denoising and super-resolution. *arXiv preprint arXiv:1905.02538*, 2019. 2

[59] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12179–12188, 2021. 1, 2

[60] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 2020. 1, 6

[61] Daniel Scharstein and Richard Szeliski. High-accuracy stereo depth maps using structured light. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, volume 1, pages I–I. IEEE, 2003. 2

[62] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 2, 3, 4

[63] Max Schwarz, Anton Milan, Arul Selvam Periyasamy, and Sven Behnke. Rgb-d object detection and semantic segmentation for autonomous manipulation in clutter. *The International Journal of Robotics Research*, 37(4-5):437–451, 2018. 1

[64] Ana Serrano, Incheol Kim, Zhili Chen, Stephen DiVerdi, Diego Gutierrez, Aaron Hertzmann, and Belen Masia. Motion parallax for 360 rgbd video. *IEEE Transactions on Visualization and Computer Graphics*, 25(5):1817–1827, 2019. 1

[65] Sudipta N Sinha, Philippos Mordohai, and Marc Pollefeys. Multi-view stereo via graph cuts on the dual of an adaptive tetrahedral mesh. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE, 2007. 2

[66] Matthew Tancik, Vincent Casser, Xinchen Yan, Sabeek Pradhan, Ben Mildenhall, Pratul P Srinivasan, Jonathan T Barron, and Henrik Kretzschmar. Block-nerf: Scalable large scene neural view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8248–8258, 2022. 1

[67] Vladimir Tankovich, Christian Hane, Yinda Zhang, Adarsh Kowdle, Sean Fanello, and Sofien Bouaziz. Hitnet: Hierarchical iterative tile refinement network for real-time stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14362–14372, 2021. 1, 2

[68] Bill Triggs, Philip F McLauchlan, Richard I Hartley, and Andrew W Fitzgibbon. Bundle adjustment—a modern synthesis. In *International workshop on vision algorithms*, pages 298–372. Springer, 1999. 4

[69] Benjamin Ummenhofer, Huizhong Zhou, Jonas Uhrig, Nikolaus Mayer, Eddy Ilg, Alexey Dosovitskiy, and Thomas Brox. Demon: Depth and motion network for learning monocular stereo. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5038–5047, 2017. 1

[70] Benjamin Ummenhofer, Huizhong Zhou, Jonas Uhrig, Nikolaus Mayer, Eddy Ilg, Alexey Dosovitskiy, and Thomas Brox. DeMoN: Depth and motion network for learning monocular stereo. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017. 7

[71] Neal Wadhwa, Rahul Garg, David E Jacobs, Bryan E Feldman, Nori Kanazawa, Robert Carroll, Yair Movshovitz-Attias, Jonathan T Barron, Yael Pritch, and Marc Levoy. Synthetic depth-of-field with a single-camera mobile phone. *ACM Transactions on Graphics (ToG)*, 37(4):1–13, 2018. 1

[72] Lijun Wang, Xiaohui Shen, Jianming Zhang, Oliver Wang, Zhe Lin, Chih-Yao Hsieh, Sarah Kong, and Huchuan Lu. Deeplens: shallow depth of field from a single image. *arXiv preprint arXiv:1810.08100*, 2018. 1

[73] Zirui Wang, Shangzhe Wu, Weidi Xie, Min Chen, and Victor Adrian Prisacariu. Nerf–: Neural radiance fields without known camera parameters. *arXiv preprint arXiv:2102.07064*, 2021. 3

[74] Mial E Warren, David Podva, Preethi Dacha, Matthew K Block, Christopher J Helms, John Maynard, and Richard F Carson. Low-divergence high-power vcsel arrays for lidar application. In *Vertical-Cavity Surface-Emitting Lasers XXII*, volume 10552, page 105520E. International Society for Optics and Photonics, 2018. 2

[75] Tsachy Weissman, Erik Ordentlich, Gadiel Seroussi, Sergio Verdú, and Marcelo J Weinberger. Universal discrete denoising: Known channel. *IEEE Transactions on Information Theory*, 51(1):5–28, 2005. 3

[76] Bartlomiej Wronski, Ignacio Garcia-Dorado, Manfred Ernst, Damien Kelly, Michael Krainin, Chia-Kai Liang, Marc Levoy, and Peyman Milanfar. Handheld multi-frame super-resolution. *ACM Transactions on Graphics (TOG)*, 38(4):1–18, 2019. 2, 3

[77] Ying Xiong, Ayan Chakrabarti, Ronen Basri, Steven J Gortler, David W Jacobs, and Todd Zickler. From shading to local shape. *IEEE transactions on pattern analysis and machine intelligence*, 37(1):67–79, 2014. 2

[78] Yalin Xiong and Steven A Shafer. Depth from focusing and defocusing. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 68–73. IEEE, 1993. 2

[79] Hao-Hsiang Yang, Wei-Ting Chen, and Sy-Yen Kuo. S3net: A single stream structure for depth guided image relighting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 276–283, 2021. 1

[80] Jinyu Yang, Zhe Li, Song Yan, Feng Zheng, Aleš Leonardis, Joni-Kristian Kämäräinen, and Ling Shao. Rgbd object tracking: An in-depth review. *arXiv preprint arXiv:2203.14134*, 2022. 1

[81] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4578–4587, 2021. 1

[82] Fisher Yu and David Gallup. 3d reconstruction from accidental motion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3986–3993, 2014. 2, 3, 5

[83] Song Zhang. High-speed 3d shape measurement with structured light methods: A review. *Optics and Lasers in Engineering*, 106:119–131, 2018. 2

[84] Michael Zollhöfer, Patrick Stotko, Andreas Görlitz, Christian Theobalt, Matthias Nießner, Reinhard Klein, and Andreas Kolb. State of the art on 3d reconstruction with rgb-d cameras. In *Computer graphics forum*, volume 37, pages 625–652. Wiley Online Library, 2018. 1