Incorporating dense metric depth into neural 3D representations for view synthesis and relighting

Arkadeep Narayan Chaudhury[†,⋆], Igor Vasiljevic[⋆], Sergey Zakharov[⋆], Vitor Guizilini[⋆], Rares Ambrus[⋆], Srinivasa Narasimhan[†], and Christopher G. Atkeson[†]
[†]The Robotics Institute, Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, PA 15213, USA
[⋆]Toyota Research Institute
4440 El Camino Real, Los Altos, CA 94022, USA
arkadeec@alumni.cmu.edu

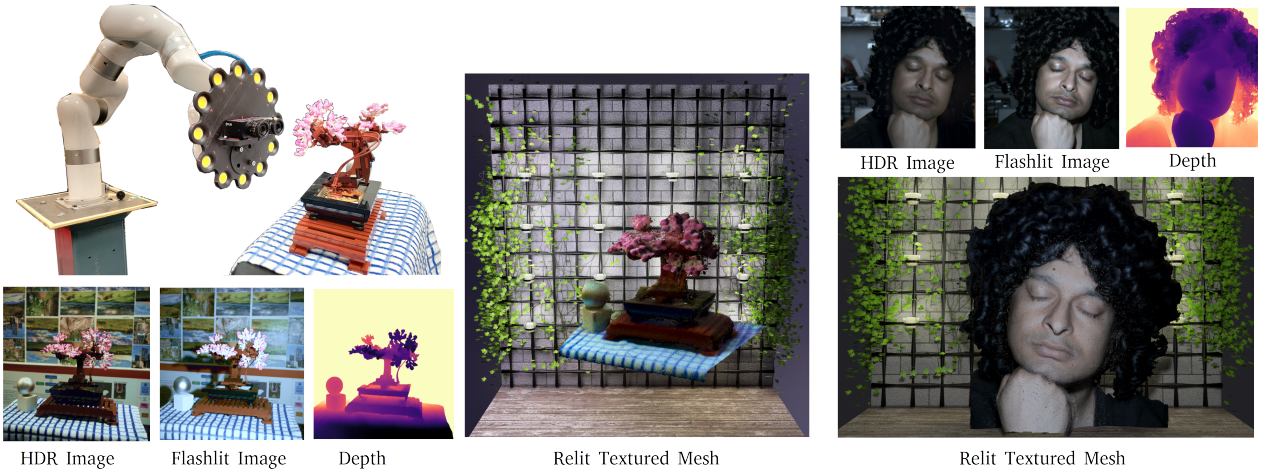


Figure 1. We present an approach for the photo-realistic capture of small scenes by incorporating dense metric depth, multi-view, and multi-illumination images into neural 3D scene understanding pipelines. We use a robot mounted multi-flash stereo camera system, developed in-house, to capture the necessary supervision signals needed to optimize our representation with a few input views. The reconstruction of the LEGO plant and the face were generated with 11 and 2 stereo pairs respectively. We relight the textured meshes using [11]. Background design by [42].

1. Introduction

Capturing photo realistic appearance and geometry of scenes is a fundamental problem in computer vision and graphics with a set of mature tools and solutions for content creation [12, 67], large scale scene mapping [5], augmented reality and cinematography [6, 80, 97]. Enthusiast level 3D photogrammetry, especially for small or tabletop scenes, has been supercharged by more capable smartphone cameras and new toolboxes like RealityCapture and NeRFStudio. A subset of these solutions are geared towards view synthesis where the focus is on photo-realistic view interpolation rather than recovery of accurate scene geometry. These solutions take the "shape-radiance ambiguity"[58] into stride by decoupling the scene transmissivity (related to geometry) from the scene appearance prediction. But without diverse training views, several neural scene representations (e.g. [39, 73, 76]) are prone to poor shape reconstructions while estimating accurate appearance.

By only reasoning about appearance as cumulative radiance weighted with the scene's transmissivity, one can achieve convincing view interpolation results, with the quality of estimated scene geometry improving with the diversity and number of training views. However, capturing a diverse set of views, especially for small scenes, often becomes challenging due to the scenes' arrangement.

Without dense metric depth measurements, researchers have used sparse depth from structure from motion [30, 53, 83], and dense monocular depth priors [108] to improve reconstruction, with a focus on appearance. Assimilating dense non-metric depth (e.g. [20, 33]) is often challenging due to the presence of an unknown affine degree of freedom which needs to be estimated across many views.

However, without diversity of viewpoints, measuring the geometry directly is often useful. Several hardware solutions for digitizing objects exist, ranging from consumer level 3D scanners (e.g. [91]), and room scale metrology devices [2, 3, 69] to high precision hand held 3D scanners (e.g. [1]). Although these systems measure geometry very accurately, they interpret appearance as diffuse reflectance and often fall short in modelling view-dependent effects.

Despite the known effectiveness of incorporating depth and widespread availability of dense metric depth sensors in smartphone cameras [110, 112] and as standalone devices [4, 54], incorporation of dense metric depth into neural 3D scene understanding is underexplored. In this work:

1. we present a method to incorporate dense metric depth into the training of neural 3D fields, enabling state-of-the-art methods to use dense metric depth with minor changes.
2. We investigate an artifact (Fig. 4) commonly observed while jointly refining shape and appearance. We identify its cause as existing methods' inability to differentiate between depth and texture discontinuities. We address it by using depth edges as an additional supervision signal.

We demonstrate our ideas using a robot mounted multi-flash stereo camera rig developed in-house from off-the-shelf components. This device allows us to capture a diverse range of scenes with varying complexity in both appearance and geometry. Using the captured data, we demonstrate results in reconstruction, view interpolation, geometry capture, and relighting with a few views. We hope that our full-stack solution comprising of the camera system and algorithms will serve as a test bench for automatically capturing small scenes in the future. Additional results may be viewed at https://stereomfc.github.io and in the supplementary document.

2. Related Work

View synthesis and reconstruction of shapes from multiple 3D measurements is an important problem in computer vision with highly efficient and general solutions like volumetric fusion [26], screened Poisson surface reconstruction [52], patch based dense stereopsis [41] and joint refinement of surface and appearance [27]. While these continue to serve as robust foundations, they fall short in capturing view-dependent appearance. Additionally, even with arbitrary levels of discretization, they often oversmooth texture and surfaces due to data association relying on weighted averages along the object surface.

Recent neural 3D scene understanding approaches (e.g. [62, 100, 105]) have avoided this by adopting a continuous implicit volumetric representation to serve as the geometric and appearance back-end of the view synthesizer. Together with continuous models, reasoning about appearance as radiance, and high frequency preserving embeddings[96], these approaches serve as highly capable view interpolators by reliably preserving view dependent appearance and minute geometric details. More recent work has included additional geometric priors in the form of monocular depth supervision [108], sparse depth supervision from structure-from-motion toolboxes [95], dense depth maps [9, 84], patch based multi-view consistency [40], and multi-view photometric consistency under assumed surface reflectance functions [46]. Our work builds on the insights from using dense depth supervision to improve scene understanding with only a few training views available.

Novel hardware is often used for collecting supervision signals in addition to color images to aid 3D scene understanding. [8] demonstrate a method to incorporate a time-of-flight sensor. [89] demonstrate a method to extract geometric and radiometric cues from scenes captured with a commercial RGBD sensor and improve view synthesis with a few views. Event based sensors have also been used to understand poorly lit scenes with fast moving cameras [56, 66]. Researchers have also combined illumination sources with cameras to capture photometric and geometric cues for dense 3D reconstruction of scenes with known reflectances [17, 44]. Similarly, [7, 21, 86, 87] capture geometry and reflectance of objects by refining multi-view color, depth and multi-illumination images. Given the recent advances in stereo matching [103] we use a stereo camera to collect data for view synthesis to disambiguate between shape and appearance at capture.

Pairing illumination sources with imaging can improve reasoning about the appearance in terms of surface reflectance parameters. [50, 87, 114] approaches the problem of material capture using a variety of neural and classical techniques. [10, 21, 51, 111] leverage recent neural scene understanding techniques to jointly learn shape and appearance as reflectance of the scene. Our work also pairs illumination sources with stereo cameras to capture multi-illumination images from the scene and we build on modern neural techniques for view synthesis and relighting the scene.

3. Method

We follow related works [62, 100, 101, 105] and represent the scene with two neural networks – an intrinsic network $\mathcal{N}(\theta)$ and an appearance network $\mathcal{A}(\phi)$ which are jointly optimized to capture the shape and appearance of the object. $\mathcal{N}(\theta)$ is a multi-layer perceptron (MLP) with parameters (θ) and uses multi-level hash grids to encode the inputs [62]. It is trained to approximate the intrinsic properties of the scene – the scene geometry as a neural signed distance field $\mathcal{S}(\theta)$ and an embedding $\mathcal{E}(\theta)$. The appearance network $\mathcal{A}(\phi)$ is another MLP which takes $\mathcal{E}(\theta)$ and a frequency encoded representation of the viewing direction and returns the scene radiance along a ray.

Prior work has jointly learned $\mathcal{S}, \mathcal{N}, \mathcal{A}$ with only multi-view images by optimizing a loss in the form of Eq. (7)

(a) (b) (c) (d) (e)

Figure 2. A snapshot of the important supervision signals. We capture a high dynamic range image [71] and display it after tonemapping [81] in Fig. 2a. Figure 2b shows the scene depth (in mm) from stereo. Figure 2c displays the likelihood of each pixel falling on a depth edge. Figure 2d shows the object surface normals. We note that unlike conventional stereo matching [47], [103] returns locally smooth surfaces and often ignores local texture variations but is less noisy. The inset shows the surface normals on the textured aluminum plate calculated as gradients of depth from conventional stereo matching. Finally, Fig. 2e identifies the pixels with the largest appearance variation due to moving lights. We used the system in Fig. 5 to capture the data.

using stochastic gradient descent [55] along a batch of rays projected from known camera centers to the scene

$$\ell = \ell_C + \lambda_g \ell_D + \lambda_c \mathbb{E}(|\nabla^2_{\mathbf{x}} \mathcal{S}(\mathbf{x}_s)|), \tag{1}$$

where λs are hyperparameters and the third term in Eq. (7) is the mean surface curvature minimized against the captured surface normals (see [62]). As the gradients of the loss functions ℓ_C (appearance loss) and ℓ_D (geometry loss) propagate through $\mathcal{A}$ and $\mathcal{N}$ (and $\mathcal{S}$ as it is part of $\mathcal{N}$) the appearance and geometry are learned together.

We describe our method of incorporating dense metric depth in Sec. 3.1 which enables a variety of neural 3D representations (Sec. 3.3) to use it. In Sec. 3.2 we jointly optimize shape and appearance of a scene using information about scene depth edges.

3.1. Incorporating dense metric depth

Given a large number of orthogonal view pairs (viewpoint diversity), and the absence of very strong view dependent effects, Eq. (7) is expected to guide $\mathcal{S}$ to towards an unbiased estimate of the true scene depth (see e.g. [43]). We can accelerate the convergence by providing high quality biased estimate of the scene depth. Given the quality of modern deep stereo [103] and a well calibrated camera system, a handful of aligned RGBD sequences can serve as a good initial estimate of the true surface depth in absence of diverse viewpoints.

In this section we describe our method to directly optimize $\mathcal{S}$ with estimates of true surface depth to any surface point $\mathbf{x}_s$. Although [27, 77, 115] use the depth estimates directly, they fall short of modelling view-dependent effects. To avoid that, we elect to learn a continuous and locally smooth function that approximates the signed distance function of the surface $\mathbf{x}_s$ which can then be transformed to scene density [77, 100, 105]. To do this, we roughly follow [45] and consider a loss function of the form

$$\ell_D(\theta) = \ell_{\mathbf{x}_s} + \lambda \mathbb{E}(||\nabla_{\mathbf{x}} \mathcal{S}(\mathbf{x}^{\Delta}, \theta)|| - 1)^2 \tag{2}$$

$$\text{where, } \ell_{\mathbf{x}_s} = \frac{1}{N} \Sigma_{\forall \mathbf{x}} \left[\mathcal{S}(\mathbf{x}, \theta) + 1 - \langle \nabla_{\mathbf{x}} \mathcal{S}(\mathbf{x}, \theta), \mathbf{n}_x \rangle \right].$$

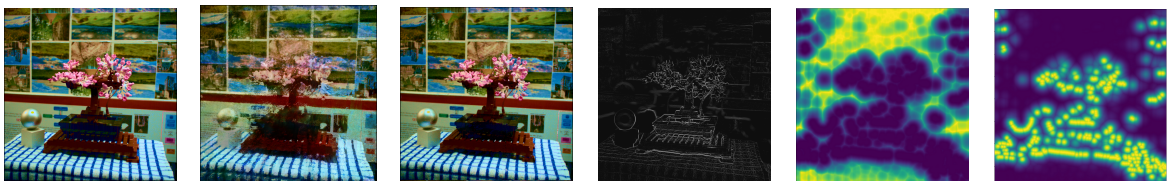
Through the two components of $\ell_{\mathbf{x}_s}$, the loss encourages the function $\mathcal{S}(\mathbf{x}, \theta)$ to vanish at the observed surface points and the gradients of the surface to align at the measured surface normals $(\mathbf{n}_x)$. The second component in Eq. (2) is the Eikonal term [24] which encourages the gradients of $\mathcal{S}$ to have a unit L_2 norm everywhere. The individual terms of Eq. (2) are averaged across all samples in a batch corresponding to N rays projected from a known camera.

The Eikonal constraint applies to the neighborhood points $\mathbf{x}_s^{\Delta}$ of each point in $\mathbf{x}_s$. [45] identifies candidate $\mathbf{x}_s^{\Delta}$ through a nearest neighbor search, whereas [105] identifies $\mathbf{x}^{\Delta}$ through random perturbations of the estimated surface point along the projected ray. As we have access to depth maps, we identify the variance of the neighborhood of $\mathbf{x}_s$ through a sliding window maximum filter on the depth images. This lets us avoid expensive nearest neighbor lookups for a batch of $\mathbf{x}_s$ to generate better estimates of $\mathbf{x}_s^{\Delta}$ than [105] at train time. As a result, convergence is accelerated – ($\sim 100\times$ over [45]) with no loss of accuracy. As we used metric depth, noisy depth estimates for parts of the scene are implicitly averaged by $\mathcal{S}$ optimized by minimizing Eq. (2), making us more robust to errors than [108]. We provide more details in the supplementary material.

3.2. Incorporating depth edges in joint optimization of appearance and geometry

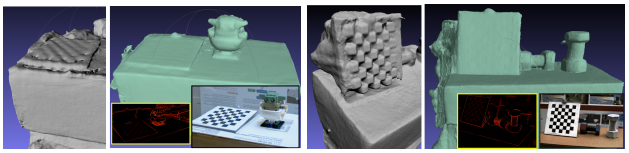
Prior works [13, 15, 27, 48, 62, 100, 105, 115] show the benefits of jointly refining geometry and appearance as it affords some degree of geometric super-resolution and more stable training. However, some pathological cases may arise when the scene has a large variation in appearance corresponding to a minimal variation in geometry across two neighboring surface points – $\mathbf{x}_s$ and $\mathbf{x}_s^{\Delta}$. We investigate this effect by considering an extreme case – a checkerboard printed on matte paper with an inkjet printer, where there is no geometric variation (planar geometry) or view dependent artifacts (ink on matte paper is close to Lambertian) corresponding to a maximum variation in appearance (white on black). The qualitative results are presented in Fig. 4.

Consider two rays $\vec{r}_{\mathbf{x}_s}$ and $\vec{r}_{\mathbf{x}_s^{\Delta}}$ connecting the camera center and two neighboring points $\mathbf{x}_s$ and $\mathbf{x}_s^{\Delta}$ on two sides



(a) (b) (c) (d) (e) (f)

Figure 3. Overview of our sampling process during training. Figure 3a is the ground truth test image. Figure 3b is the reconstruction of the test image after training has progressed 15% (15k gradient steps), Fig. 3c is the reconstruction of the test image at the end of training (100k gradient steps). Figure 3d denotes the per-pixel likelihoods of depth edges in the scene at the same view captured with our device. We note in Fig. 3b, the parts of scene with complicated geometry (foliage with many depth edges) have lower fidelity of appearance in the reconstruction at an earlier stage of training, which gradually improves in Fig. 3c. Figure 3e indicates the per-pixel sampling likelihood *if the test view were to be used for training*, at a training progress of 10%, Fig. 3f indicates the same at a progress of 90%. Equation (3) is used to draw the samples: $\alpha = 0.1$ and 0.9 respectively for Figs. 3e and 3f. Brighter color indicates higher sampling likelihood.



(a) (b)

Figure 4. We demonstrate a corner case of jointly refining appearance and geometry. The left insets of Figs. 4a and 4b are the scene geometries recovered in the worst cases, the right insets display the better meshes recovered using the method described in Sec. 3.2. An image used for training and the edge map used for sampling are in the insets. We recommend zooming into the figure for details. Corresponding quantitative results are in Tab. 4.

of an checkerboard edge included in the same batch of the gradient descent. The total losses for those rays depend on the sum of the geometry and appearance losses (Eq. (7)). By default, the current state of the art [62, 101, 106] etc. do not have a mechanism to disambiguate between texture and geometric edges (depth discontinuities).

As seen in Fig. 4, given unsuitable hyperparameters, the approaches will continue to jointly update both geometry and appearance to minimize a combined loss (Eq. (7)). This can often result in pathological reconstructions (left insets in Fig. 4) due to ℓ_C gradients dominating over ℓ_D. By gradually increasing the modelling capacity of $\mathcal{N}$ we can somewhat avoid this artifact and force the gradient updates to focus on $\mathcal{A}$ to minimize the cumulative loss. [62] recognize this and provide an excellent set of hyperparameters and training curricula to gradually increase the modelling capacity of $\mathcal{N}(\phi)$. This results in remarkable geometric reconstructions for well known datasets [49, 57]. Alternatively, if we have per-pixel labels of geometric edges ($\mathbf{E}$, Figs. 2c and 3d), we can preferentially sample image patches with low variation of geometric features when the model capacity is lower ($\mathcal{S}(\theta)$ tends to represent smoother surfaces), and focus on image patches with geometric edges when the model capacity has increased. The modelling capacity of $\mathcal{A}(\phi)$ never changes.

Figure 3 describes our sampling procedure while learning a scene with a variety of geometric and texture edges. Equation (3) is used to draw pixel samples – the probability of drawing pixel p_i is calculated as a linear blend of the likelihood that it belongs to the set of edge pixels $\mathbf{E}$ and α is a scalar ($\alpha \in [0, 1]$) proportional to the progress of the training.

$$P(p_i|\alpha) = (1 - \alpha)P(p_i \in \mathbf{E}) + \alpha P(p_i \notin \mathbf{E}) \quad (3)$$

To preserve the geometric nature of the edges while ruling out high frequency pixel labels, we use Euclidean distance transform [34] to dilate $\mathbf{E}$ before applying Eq. (3). We provide implementation details in the supplementary material for reproducibility. We discuss quantitative results in Sec. 5.2.

3.3. Baselines augmented with depth

As baselines, we augment four state-of-the-art methods to incorporate metric depth:

AdaShell^{++} is our implementation of AdaptiveShells [101] using metric depth. We retain the formulations for the scene geometry and appearance models, and adapt the formulation of the "shells" to use dense metric depth. Through AdaShell^{++}we also demonstrate how dense metric depth can combine the advantages of volumetric and surface based representations in Fig. 7.

VolSDF^{++}is our augmented version of [105, 106], where we use the metric depths along the rays to optimize the geometry Eq. (2). All the other parts of the original approaches, including the methods for generating samples to minimize Eq. (7) and scene density transforms are left unchanged.

NeUS^{++}is our augmented version of [62], where we also use Eq. (2) to optimize the geometry. The rest of the algorithm including the background radiance field is left intact.

UniSurf^{++}is our deliberately hamstrung version of [77] where we force the samples generated for the volumetric rendering step to have a very low variance around the current biased estimate of the surface. This makes the algorithm necessarily indifferent to the relative magnitudes of

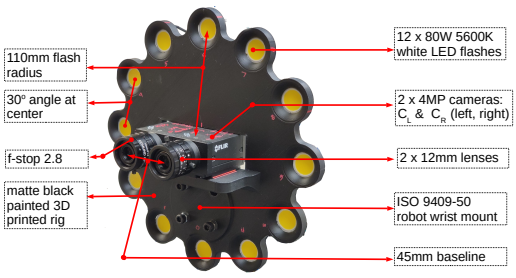


Figure 5. A multi-flash stereo camera to image small scenes.

ℓ_D and ℓ_C in Eq. (7), and helps us exaggerate the pathological effects of not segregating texture and geometric edges. We choose to name the method UniSurf^{++}after we (and [13]) observed that original method was vulnerable to this artifact under certain hyperparameter choices.

Across all the methods, we implement and train $\mathcal{N}(\theta)$ following [62], and all of them use the same appearance network $\mathcal{A}(\phi)$. AdaShell^{++}and UniSurf^{++}require a warm start – $\mathcal{S}$ pre-optimized for 5K gradient steps. All methods except UniSurf^{++}use the sampling strategy from Sec. 3.2. More details are in the supplementary material.

4. Setup and Dataset

4.1. A multi-flash stereo camera

In addition to multi-view images, scene depth and depth edges are valuable signals to train neural 3D representations. To capture all the supervision signals, we designed and fabricated a multi-flash stereo camera based on insights from [35, 79], with off-the-shelf parts. We capture data by moving our camera rig in front of objects. For each camera pose we capture a stereo pair of high dynamic range (HDR) images, two depth maps from left and right stereo, two corresponding image aligned surface normals (as gradient of depth maps). We first tonemap the HDR images [81] and in-paint them with the depth edges before using [103] to preserve intricate surface details in the depth maps. Additionally we capture 12 pairs of multi-illumination images for 12 flash lights around the cameras, one light at a time. From the multi-flash images we recover a per pixel likelihood of depth edges in the scene and a label of pixels with a large appearance variation under changing illumination – relating to the specularity. We detail the design of our rig and the capture processes in the supplementary material. Figure 2 shows a snapshot of the data captured, Fig. 5 illustrates our camera rig prototype.

We elected to calculate depth from stereo because it performs better than the following three alternatives we tested. 1) Recovering geometry from intrinsic-image-decomposition [28] and photometric stereo with a few lights [17] did not yield satisfactory results. PIE-Net[28] requires 256×256 images which were too low-resolution for reconstruction and, our captures were out-of-distribution for the pre-trained model. [17] assumes fixed lights – we'd need new light calibrations per-view. 2) Self-calibrating-photometric-stereo [18, 59], demonstrated on [90], needs 50-80 light views and accurate masks which we do not capture. Also, our lights are much closer to the camera than [60, 90]. And, 3) modern camera-projector systems [19, 75] yield better estimates of geometry than stereo, but is not fast enough to capture human subjects (Fig. 8).

	BMVS	DTU	ReNe	DGT^{+}	PDR	OIl.	Ours
Depth	✓	✓	×	✓	×	×	✓
Light	OLAT	×	OLAT	OLAT	×	OLAT	Flash
Pol.	×	×	×	×	✓	✓	×
Spec.	×	×	×	×	✓	×	✓
D.E.	×	×	×	×	×	×	✓
HDR	✓	×	×	✓	×	✓	✓
Illum.	✓	×	×	✓	✓	✓	×

Table 1. We identify some differences between our dataset and a few established datasets: BMVS[104], DTU[49], ReNe[98], DiLiGenT[90], DiLiGenT-MV[90] (both abbreviated as DGT^{+}), PaNDoRa (PDR)[29] and Open-Illumination (OIl)[64]. OLAT: one light at a time, Flash: camera flash $< 0.1f$ away from camera, Pol.: polarization information, Spec.: specularity labels, D.E.: depth edge labels, HDR: High dynamic range images, Illum.: Illumination model supplied.

4.2. Dataset

Although a dataset is not the primary contribution of our research, we capture some salient aspects of the scene that are not present in several established datasets. We identify these aspects in Tab. 1. In the rows labeled "specularity" and "depth edges" we note if the dataset has explicit labels for the specular nature of the pixel or a presence of a depth edge at that pixel respectively. Under "illum. model" we note if an explicit illumination model is present per scene – we do not capture an environment illumination model, and instead provide light poses. PaNDoRa does not have explicit specularity labels but polarization measurements at pixels may be used to derive high quality specularity labels, which are better than what our system natively captures. We differentiate between "OLAT" (one light at a time) and "flash" by the location of the source of illumination. Similar to [21], our flashes are parallel to the imaging plane, located close (∼0.1f) to the camera, as opposed to ReNE and OpenIllumination.

5. Experiments and results

5.1. Accuracy of incorporating metric depth

We reconstruct synthetic scenes with ground truth depth from [9, 84] to measure the accuracy of our technique. We use 12-15 RGBD images to reconstruct the scenes and train for an average of 30k gradient steps (∼1500 epochs) in about 75 minutes. In contrast, [9, 84] use 300+ RGBD tuples and 9+ hours of training on comparable hardware. Notably, [9] also optimizes for noise in camera poses and re-

Scene	NRGBD	BF	AdaShell^{++}	NeUS^{++}	VolSDF^{++}
greenroom	**0.013**	0.024	0.015	0.016	0.014
staircase	0.045	0.091	0.024	**0.009**	0.020
kitchen I	0.252	0.234	0.044	**0.036**	0.047
kitchen II	**0.032**	0.089	0.045	**0.032**	0.060

Table 2. Accuracy of reconstruction from un-posed RGBD images. For ***un-posed*** RGBD images, we compare the accuracy of scene reconstruction using AdaShell^{++}, NeUS^{++}, and VolSDF^{++}with NeuralRGBD (NRGBD) [9] and BundleFusion (BF) [27]. We report normalized Chamfer distances (lower is better) across four synthetic scenes from [9].

Scene	R 0	R 1	R 2	O 0	O 1	O 2	O 3	O 4
[84]	0.61	0.41	0.37	0.38	0.48	0.54	0.69	0.72
AS^{++}	**0.11**	**0.10**	**0.09**	**0.12**	**0.06**	**0.08**	**0.14**	**0.11**

Table 3. Accuracy of reconstruction from posed RGBD images. For RGBD images with ***ground-truth poses***, we compare the accuracy of reconstructing the scene between AdaShell^{++}(AS^{++}) and PointSLAM[84]. We report the mean L_1 distances in cm (lower is better) across eight synthetic scenes from the Replica Dataset [94]. The scenes with prefix R are the room scenes, scenes with prefix O are the office scenes.

Scene	stereo	VolSDF^{++}	NeUS^{++}	AdaShell^{++}	UniSurf^{++}
Fig. 4a	6.22	4.74	**2.77**	5.42	13.21
Fig. 4b	6.82	3.87	**3.68**	6.34	16.02

Table 4. Depth edges help prioritize learning of texture discontinuities over geometric ones. We report the RMS deviation from a plane (lower is better) for the reconstructed checkerboard surfaces in mm. We note that AdaShell^{++}performs slightly worse than volumetric methods VolSDF^{++}and NeUS^{++}. Except for UniSurf^{++}, all improve the quality of the surface measured with only stereo. Qualitative results are shown in Fig. 4.

Metric	VolSDF^{++}			NeUS^{++}			AdaShell^{++}		
27.5+	21.3	33.1	40.0	70.5	100+	100+	22.6	23.2	94.6
100K	30.28	30.33	30.69	27.82	29.45	25.31	31.56	31.45	28.27

Table 5. Training performance for view synthesis. We report two metrics - number of steps required to reach or exceed a PSNR of 27.5 and PSNR at the end of 100K gradient steps. We observe that all the baselines perform competitively and ignoring depth and additional supervision signals (last two columns) leads to failures in the view synthesis tasks.

ports metrics with ground truth and optimized poses. We report the best metric among these two. [27] registers the images themselves. We register the RGBD images with a combination of rigid and photometric registration [78, 85, 113]. We present the quantitative results in Tabs. 2 and 3. We replicate or out-perform the baselines by using a fraction of the training data and gradient steps. Among all methods discussed in Sec. 3.3, AdaShell^{++}and VolSDF^{++}demonstrate similar performance, NeUS^{++}recovers a smoother surface at the expense of $\sim$ 1.25$\times$ more gradient steps. Our errors on these synthetic datasets closely reflect the performance of [45] on approximating surfaces from low noise point clouds. These datasets do not have large view dependent appearance variations to affect the gradient updates.

5.2. The effect of depth edges in training

We tested fused RGBD maps from stereo and four baselines from Sec. 3.3 to investigate the effect of depth and texture edges. We use edge guided sampling (Sec. 3.2 and Eq. (3)) for all except stereo and UniSurf^{++}to prioritize learning geometric discontinuities over appearance. We present the results in Fig. 4 and Tab. 4. All of the baselines except UniSurf^{++}improve the reconstruction accuracy due to segregation of texture and depth edges. The smoothness enforced by the curvature loss in Eq. (7) also improves the surface reconstruction over stereo.

5.3. View synthesis with dense depth

Incorporating dense metric depth and our sampling strategy from Secs. 3.1 and 3.2 enables AdaShell^{++}, VolSDF^{++}, and NeUS^{++}to perform competitively across challenging scenes. Scene A (Fig. 6(a)) looks at a couple of reflective objects with large variation in view dependent appearance. Additionally, there are large local errors in the captured depth maps due to specularities in the scene. We capture six stereo pairs, train on 11 images and test on one image. Scene B (Fig. 6(b)) features a rough metallic object of relatively simple geometry captured by a 16mm lens (450 mm focal length, shallow depth of field). We capture four stereo pairs, train on seven images and test on one image. Scene C (Fig. 6(c)) features a fairly complicated geometry and is captured with 12 stereo pairs. We train on 22 images and test on two. Quantitative results of our experiments are in Tab. 5. We observe that AdaShell^{++}, which is roughly 15% faster per gradient step than VolSDF^{++}, generally converges the fastest (wall clock time) to a target PSNR. When the geometry is very complicated (scene C), an equally complicated sampling volume negates the efficiency gains of our sampler. We could not find good parameters for UniSurf^{++}for any of these sequences.

View synthesis was unsuccessful without the inclusion of dense depth. We trained VolSDF^{++}with no depth supervision (equivalent to [105]) until saturation (less than 0.1 PSNR increase for 1000 consecutive epochs). The reconstructions, none of which had a PSNR of 18 or higher, are shown in the last column of Fig. 6.

5.4. Using noisy depth

To investigate the effects of noise in the depth maps, we obtain the depths of scenes using conventional stereo. We used semi-global matching stereo [47] with a dense census cost [109] and sub-pixel refinement on tone mapped HDR images to calculate the surface depth. Surface normals were calculated using the spatial gradients of the depth maps. To focus on the performance of our approaches, we did not filter or smooth the depth obtained from conventional stereo. From the top of Tab. 6, we observe that NeUS^{++}strictly improves the quality of the surface reconstructed from just

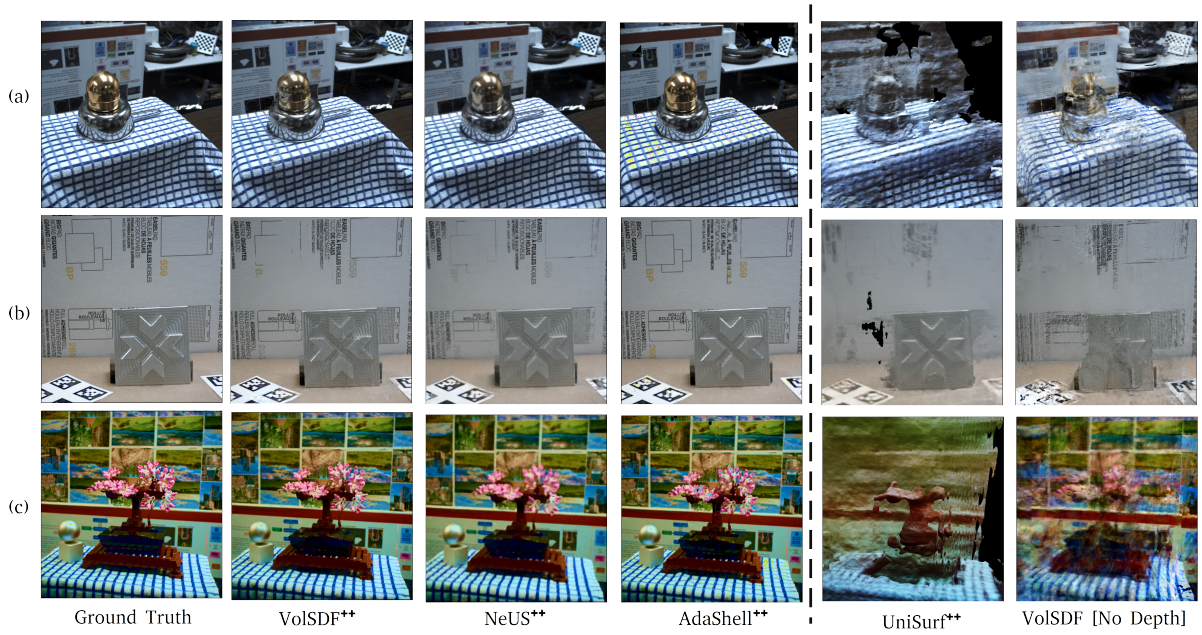


Figure 6. Relative performance of the baselines. Quantitative results in Tab. 5, discussions in Sec. 3.3.

noisy stereo (row 1 and 2 versus row 3), especially when edge sampling is enabled. If the end goal is just view synthesis, AdaShell^{++}, which blends the advantages of volumetric and surface based rendering, performs equally well with large noise in depth, whereas NeUS^{++}takes many more iterations to converge. This indicates that photorealistic view synthesis with a volumetric renderer is possible with noisy depth data. However conventional stereo often introduces large local errors which our approaches were unable to improve significantly.

In the presence of noisy depth, the quality of the reconstructed surface was enhanced through edge-based sampling (Sec. 3.2 and Eq. (3)). Our sampling strategy allocated samples away from depth edges, where the noise was more prevalent, leading to fewer gradient steps spent modelling areas with higher noise. Table 6 presents the quantitative details of the experiment.

scene	Fig. 7(a)[5]	Fig. 6(b)[7]	Fig. 4b[5]
edge sampling	**491**	**403**	**225**
no edge sampling	593	419	251
noisy stereo	600	523	369
27.5+ AdaShell^{++}w/ noise	7.93	20.2	5.12
27.5+ AdaShell^{++}w/o noise	7.85	23.2	2.71
25.0+ NeUS^{++}w/ noise	49.4	36.0	32.9

Table 6. Effect of noisy depth and depth edges. Top: The surface reconstruction quality (Hausdorff distance, lower is better) with conventional (noisy) stereo compared with surface recovered by NeUS^{++}on learned stereo. Bottom: gradient steps (in 1000s lower is faster) required to surpass a test time target PSNR. We specify the count of training views in [ ] braces.

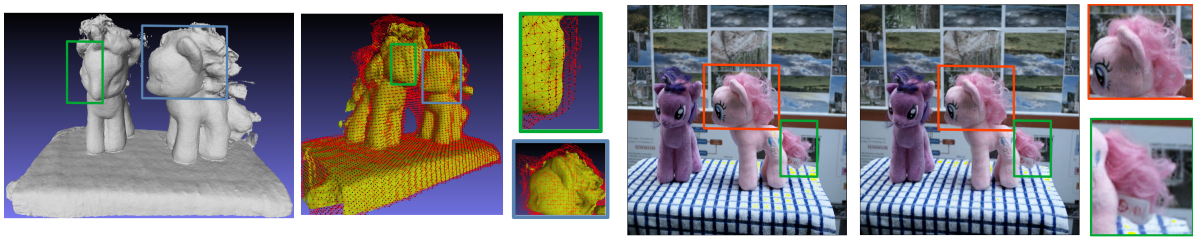
Scene	F_1	F_2	F_3	F_4	S_1	S_2
mask	**28.95**	**31.19**	**29.56**	**27.28**	**25.18**	**24.51**
no mask	27.17	30.05	27.82	25.46	23.65	21.88

Table 7. Relighting scenes with a volumetric renderer. We report PSNR (higher is better) under two heads – masked and unmasked relit images, to offset the effects of incorrect shadows cast on the background. The unmasked reconstructions generally have a poorer PSNR because our implicit scene understanding approach does not approximate a ray tracer and cannot cast correct shadows on the background. The lower PSNR for reconstructing the shiny objects is mainly due to the inability of the network to model saturation caused by reflection. Results in Fig. 8.

5.5. Relighting

We capture multi-illumination images with known light poses and recover geometry independently of appearance. This allows us to infer the illumination dependent appearance using a combination of physically based appearance parameters – e.g. the Principled BRDF[14]. As a benchmark, we upgraded the closest related work, [21], which uses the full gamut of the Disney BRDF parameters, with NeUS^{++}, to incorporate dense depth. For the data we collected, the optimization process as implemented by [21], was quite brittle and some parameters (e.g. 'clearcoat-gloss') would often take precedence over other appearance parameters (e.g. 'specular-tint') and drive the optimization to a poor local minima. We demonstrate this problem in detail in the supplementary material. We found the optimization of a subset of appearance parameters ('base-color', 'specular-tint', and 'roughness') to be the most stable. [13, 111] conclude the same.

For relighting, we explore two avenues – the inference



(a) (b) (c) (d)

Figure 7. AdaShell^{++}recovers sampling volumes similar to [101]. Fig. a shows the geometry recovered with 5 RGBD tuples. Figures b displays the sampling volumes around the geometry after AdaShell^{++}has converged – we note the similarity of this step with [101]. Figs. c and d are the ground-truth and reconstructed test images. AdaShell^{++}combines the advantages of volumetric rendering (see insets in fig. d) and surface based rendering (fig. b). More details are in the supplementary materials.

(a) Face 1 (`F_1`) **(b)** Face 2 (`F_2`) **(c)** Face 3 (`F_3`) **(d)** Face 4(`F_4`) **(e)** Shiny 1 (`S_1`) **(f)** Shiny 2 (`S_2`)

Figure 8. Relighting scenes can be achieved with AdaShell^{++}trained with multi-illumination images. Discussion and results in Sec. 5.5 and Tab. 7. We capture 12 flash lit images for all the camera views and we use alternate flashes for all the training views (6 per view). Figures above show one (of six) flash configurations for the test view. More results on the project website.

step of our approach as a volumetric renderer and a mesh created with the appearance parameters as texture (Fig. 1). Quantitative and qualitative results of the volumetric renderer are shown in Tab. 7 and Fig. 8. We used [11] to unwrap the geometry and generate texture coordinates whose quality exceeded [107] and our implementation of [93]. None of our approaches worked on the ReNe dataset (Tab. 1, [98]) due to low view diversity, and the absence of metric depth. We used the labels in Fig. 2e to allocate more gradient steps for learning the regions with higher appearance variation. We provide more details in the supplementary material.

6. Limitations

Although we achieve state of the art results in view-synthesis and relighting with a few views, our approach struggles to represent transparent objects and accurately capture the geometry of reflective surfaces. [65] address the problem of reflective objects by modelling background reflections and is based on the architecture proposed by [100]. As NeUS^{++}enables [100] to use possibly noisy metric depth, it can potentially be extended to model reflective objects.

Our approaches require metric depth and depth edges for the best performance. Our approach relies on capture devices with reasonable quality depth measurements. Future work will address incorporation of monocular and sparse depth priors with depth edges.

Incorporation of metric depth introduces a strong bias, often limiting super resolution of geometry sometimes achieved in neural 3D scene representation (see e.g. [62]). Decreasing the effect of Eq. (2) during training may potentially encourage geometric superresolution and is future work.

Finally, modern grid based representations (see e.g. [31, 82]) produce very compelling view interpolation results at a fraction of the computational cost of a state of the art volumetric renderer (e.g. [76, 101]). However, they need to be "distilled" from a pre-trained volumetric view interpolator. Future work can investigate the use of depth priors to train a grid based representation directly from color and depth images.

7. Conclusions

We present a solution to incorporate dense metric depth into neural 3D reconstruction which enables state of the art geometry reconstruction. We examine a corner case of jointly learning appearance and geometry and address it by incorporating additional supervision signals. Additionally, we describe a variant of the multi-flash camera to capture the salient supervision signals needed to improve photorealistic 3D reconstruction and demonstrate a pipeline for view synthesis and relighting of small scenes with a handful of training views.

References

[1] Artec Leo, 2023. 2

[2] Ensenso XR series scanners, 2023. 2

[3] Photoneo PhoXi3D scanners, 2023. 2

[4] Azure Kinect 3D sensors, 2023. 2

[5] 3DZephyr. 3df zephyr - photogrammetry software - 3d models from photos, 2022. 1

[6] AliceVision. Alicevision meshroom, 2022. 1

[7] Louis-Philippe Asselin, Denis Laurendeau, and Jean-Francois Lalonde. Deep svbrdf estimation on real materials. In *2020 International Conference on 3D Vision (3DV)*, pages 1157–1166. IEEE, 2020. 2

[8] Benjamin Attal, Eliot Laidlaw, Aaron Gokaslan, Changil Kim, Christian Richardt, James Tompkin, and Matthew O'Toole. Törf: Time-of-flight radiance fields for dynamic scene view synthesis. *Advances in neural information processing systems*, 34:26289–26301, 2021. 2

[9] Dejan Azinović, Ricardo Martin-Brualla, Dan B Goldman, Matthias Nießner, and Justus Thies. Neural rgb-d surface reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6290–6301, 2022. 2, 5, 6, 15

[10] Sai Bi, Zexiang Xu, Pratul Srinivasan, Ben Mildenhall, Kalyan Sunkavalli, Miloš Hašan, Yannick Hold-Geoffroy, David Kriegman, and Ravi Ramamoorthi. Neural reflectance fields for appearance acquisition. *arXiv preprint arXiv:2008.03824*, 2020. 2

[11] Blender. *Blender - a 3D modelling and rendering package*. Blender Foundation, Blender Institute, Amsterdam, 2024. 1, 8, 17, 18

[12] Pierre Boudoin. Hyper capture: 3d object scan, 2023. 1

[13] Mohammed Brahimi, Bjoern Haefner, Tarun Yenamandra, Bastian Goldluecke, and Daniel Cremers. Supervol: Super-resolution shape and reflectance estimation in inverse volume rendering. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3139–3149, 2024. 3, 5, 7

[14] Brent Burley. Physically-based shading at disney. In *Acm Siggraph*, pages 1–7. vol. 2012, 2012. 7, 16, 17

[15] Ang Cao, Chris Rockwell, and Justin Johnson. Fwd: Real-time novel view synthesis with forward warping and depth. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15713–15724, 2022. 3

[16] Manmohan Chandraker, Jiamin Bai, and Ravi Ramamoorthi. On differential photometric reconstruction for unknown, isotropic brdfs. *IEEE transactions on pattern analysis and machine intelligence*, 35(12):2941–2955, 2012. 19

[17] Arkadeep Narayan Chaudhury, Leonid Keselman, and Christopher G Atkeson. Shape from shading for robotic manipulation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 8389–8398, 2024. 2, 5, 19

[18] Guanying Chen, Kai Han, Boxin Shi, Yasuyuki Matsushita, and Kwan-Yee K Wong. Self-calibrating deep photometric stereo networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8739–8747, 2019. 5

[19] Wenzheng Chen, Parsa Mirdehghan, Sanja Fidler, and Kiriakos N. Kutulakos. Auto-tuning structured light by optical stochastic gradient descent. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 5

[20] Xinjing Cheng, Peng Wang, and Ruigang Yang. Learning depth with convolutional spatial propagation network. *IEEE transactions on pattern analysis and machine intelligence*, 42(10):2361–2379, 2019. 1, 16

[21] Ziang Cheng, Junxuan Li, and Hongdong Li. Wildlight: In-the-wild inverse rendering with a flashlight. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4305–4314, 2023. 2, 5, 7, 16, 17, 18

[22] Sungjoon Choi, Qian-Yi Zhou, and Vladlen Koltun. Robust reconstruction of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5556–5565, 2015. 18

[23] Paolo Cignoni, Marco Callieri, Massimiliano Corsini, Matteo Dellepiane, Fabio Ganovelli, and Guido Ranzuglia. MeshLab: an Open-Source Mesh Processing Tool. In *Eurographics Italian Chapter Conference*. The Eurographics Association, 2008. 17

[24] Michael G Crandall and Pierre-Louis Lions. Viscosity solutions of hamilton-jacobi equations. *Transactions of the American mathematical society*, 277(1):1–42, 1983. 3

[25] CREE. Cree xlamp cxa2540, 2024. 18

[26] Brian Curless and Marc Levoy. A volumetric method for building complex models from range images. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 303–312, 1996. 2

[27] Angela Dai, Matthias Nießner, Michael Zollhöfer, Shahram Izadi, and Christian Theobalt. Bundlefusion: Real-time globally consistent 3d reconstruction using on-the-fly surface reintegration. *ACM Transactions on Graphics (ToG)*, 36(4):1, 2017. 2, 3, 6

[28] Partha Das, Sezer Karaoglu, and Theo Gevers. Pie-net: Photometric invariant edge guided network for intrinsic image decomposition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19790–19799, 2022. 5

[29] Akshat Dave, Yongyi Zhao, and Ashok Veeraraghavan. Pandora: Polarization-aided neural decomposition of radiance. In *European Conference on Computer Vision*, pages 538–556. Springer, 2022. 5

[30] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised NeRF: Fewer views and faster training for free. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 16

[31] Daniel Duckworth, Peter Hedman, Christian Reiser, Peter Zhizhin, Jean-François Thibert, Mario Lučić, Richard Szeliski, and Jonathan T. Barron. Smerf: Streamable memory efficient radiance fields for real-time large-scene exploration, 2023. 8

[32] EdmundOptics. C series fixed focal length lenses: Edmund Optics, 2024. 18

[33] Ainaz Eftekhar, Alexander Sax, Jitendra Malik, and Amir Zamir. Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10786–10796, 2021. 1, 15

[34] Pedro F Felzenszwalb and Daniel P Huttenlocher. Distance transforms of sampled functions. *Theory of computing*, 8 (1):415–428, 2012. 4

[35] Rogerio Feris, Ramesh Raskar, Kar-Han Tan, and Matthew Turk. Specular reflection reduction with multi-flash imaging. In *Proceedings. 17th Brazilian symposium on computer graphics and image processing*, pages 316–321. IEEE, 2004. 5, 19

[36] Rogerio Feris, Ramesh Raskar, Longbin Chen, Kar-Han Tan, and Matthew Turk. Discontinuity preserving stereo with small baseline multi-flash illumination. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, pages 412–419. IEEE, 2005. 19

[37] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 18

[38] FLIR. Grasshopper3 USB3 model: GS3-U3-41C6C, 2024. 18

[39] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5501–5510, 2022. 1

[40] Qiancheng Fu, Qingshan Xu, Yew Soon Ong, and Wenbing Tao. Geo-NeUS: Geometry-consistent neural implicit surfaces learning for multi-view reconstruction. *Advances in Neural Information Processing Systems*, 35:3403–3416, 2022. 2

[41] 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 (ICCV)*, 2015. 2

[42] Ginibird. Garden tea lights [license: Cc-0], 2021. 1

[43] Lily Goli, Cody Reading, Silvia Sellán, Alec Jacobson, and Andrea Tagliasacchi. Bayes' Rays: Uncertainty quantification in neural radiance fields. *CVPR*, 2024. 3, 14

[44] Paulo F. U. Gotardo, Tomas Simon, Yaser Sheikh, and Iain Matthews. Photogeometric scene flow for high-detail dynamic 3d reconstruction. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015. 2

[45] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. *arXiv preprint arXiv:2002.10099*, 2020. 3, 6, 15

[46] Vitor Guizilini, Igor Vasiljevic, Jiading Fang, Rares Ambrus, Sergey Zakharov, Vincent Sitzmann, and Adrien Gaidon. Delira: Self-supervised depth, light, and radiance fields. *arXiv preprint arXiv:2304.02797*, 2023. 2

[47] Heiko Hirschmuller. Accurate and efficient stereo processing by semi-global matching and mutual information. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, pages 807–814. IEEE, 2005. 3, 6, 19

[48] Yuxin Hou, Arno Solin, and Juho Kannala. Novel view synthesis via depth-guided skip connections. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3119–3128, 2021. 3

[49] Rasmus Jensen, Anders Dahl, George Vogiatzis, Engin Tola, and Henrik Aanæs. Large scale multi-view stereopsis evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 406–413, 2014. 4, 5, 18

[50] Kaizhang Kang, Cihui Xie, Chengan He, Mingqi Yi, Minyi Gu, Zimin Chen, Kun Zhou, and Hongzhi Wu. Learning efficient illumination multiplexing for joint capture of reflectance and shape. *ACM Trans. Graph.*, 38(6):165–1, 2019. 2

[51] Kaizhang Kang, Minyi Gu, Cihui Xie, Xuanda Yang, Hongzhi Wu, and Kun Zhou. Neural reflectance capture in the view-illumination domain. *IEEE Transactions on Visualization and Computer Graphics*, 29(2):1450–1462, 2021. 2

[52] Michael Kazhdan and Hugues Hoppe. Screened poisson surface reconstruction. *ACM Transactions on Graphics (ToG)*, 32(3):1–13, 2013. 2

[53] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 2023. 1

[54] Leonid Keselman, John Iselin Woodfill, Anders Grunnet-Jepsen, and Achintya Bhowmik. Intel RealSense stereoscopic depth cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–10, 2017. 2, 19

[55] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 3, 13

[56] Simon Klenk, Lukas Koestler, Davide Scaramuzza, and Daniel Cremers. E-nerf: Neural radiance fields from a moving event camera. *IEEE Robotics and Automation Letters*, 8(3):1587–1594, 2023. 2

[57] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (ToG)*, 36(4):1–13, 2017. 4

[58] Kiriakos N Kutulakos and Steven M Seitz. A theory of shape by space carving. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, pages 307–314. IEEE, 1999. 1

[59] Junxuan Li and Hongdong Li. Self-calibrating photometric stereo by neural inverse rendering. In *European Conference on Computer Vision*, pages 166–183. Springer, 2022. 5

[60] Min Li, Zhenglong Zhou, Zhe Wu, Boxin Shi, Changyu Diao, and Ping Tan. Multi-view photometric stereo: A robust solution and benchmark dataset for spatially varying isotropic materials. *IEEE Transactions on Image Processing*, 29:4159–4173, 2020. 5

[61] Ruilong Li, Matthew Tancik, and Angjoo Kanazawa. Nerfacc: A general nerf acceleration toolbox. *arXiv preprint arXiv:2210.04847*, 2022. 15

[62] Zhaoshuo Li, Thomas Müller, Alex Evans, Russell H Taylor, Mathias Unberath, Ming-Yu Liu, and Chen-Hsuan Lin. Neuralangelo: High-fidelity neural surface reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 3, 4, 5, 8, 13, 15

[63] Chao Liu, Srinivasa G Narasimhan, and Artur W Dubrawski. Near-light photometric stereo using circularly placed point light sources. In *2018 IEEE International Conference on Computational Photography (ICCP)*, pages 1–10. IEEE, 2018. 19

[64] Isabella Liu, Linghao Chen, Ziyang Fu, Liwen Wu, Haian Jin, Zhong Li, Chin Ming Ryan Wong, Yi Xu, Ravi Ramamoorthi, Zexiang Xu, et al. Openillumination: A multi-illumination dataset for inverse rendering evaluation on real objects. *arXiv preprint arXiv:2309.07921*, 2023. 5

[65] Yuan Liu, Peng Wang, Cheng Lin, Xiaoxiao Long, Jiepeng Wang, Lingjie Liu, Taku Komura, and Wenping Wang. Nero: Neural geometry and brdf reconstruction of reflective objects from multiview images. *arXiv preprint arXiv:2305.17398*, 2023. 8

[66] Weng Fei Low and Gim Hee Lee. Robust e-nerf: Nerf from sparse & noisy events under non-uniform motion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 18335–18346, 2023. 2

[67] LumaLabs. Luma ai: Ai for gorgeous 3d capture, 2023. 1

[68] Ricardo Martin-Brualla, Noha Radwan, Mehdi S. M. Sajjadi, Jonathan T. Barron, Alexey Dosovitskiy, and Daniel Duckworth. NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections. In *CVPR*, 2021. 16

[69] Matterport. Matterport: Drive results with digital twins., 2023. 2

[70] Nelson Max. Optical models for direct volume rendering. *IEEE Transactions on Visualization and Computer Graphics*, 1(2):99–108, 1995. 13

[71] Tom Mertens, Jan Kautz, and Frank Van Reeth. Exposure fusion. In *15th Pacific Conference on Computer Graphics and Applications (PG'07)*, pages 382–390. IEEE, 2007. 3, 18

[72] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 17

[73] 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. *Communications of the ACM*, 65(1):99–106, 2021. 1

[74] 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. *CVPR*, 2022. 18

[75] Parsa Mirdehghan, Maxx Wu, Wenzheng Chen, David B. Lindell, and Kiriakos N. Kutulakos. Turbosl: Dense accurate and fast 3d by neural inverse structured light. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 25067–25076, 2024. 5

[76] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)*, 41(4):1–15, 2022. 1, 8, 13

[77] Michael Oechsle, Songyou Peng, and Andreas Geiger. Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5589–5599, 2021. 3, 4, 13, 14

[78] Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Colored point cloud registration revisited. In *Proceedings of the IEEE international conference on computer vision*, pages 143–152, 2017. 6, 18

[79] Ramesh Raskar, Kar-Han Tan, Rogerio Feris, Jingyi Yu, and Matthew Turk. Non-photorealistic camera: depth edge detection and stylized rendering using multi-flash imaging. *ACM transactions on graphics (TOG)*, 23(3):679–688, 2004. 5, 18, 19

[80] RealityCapture. Realitycapture: 3d models from photos and/or laser scans, 2022. 1

[81] Erik Reinhard, Michael Stark, Peter Shirley, and James Ferwerda. Photographic tone reproduction for digital images. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 661–670. 2023. 3, 5, 18

[82] Christian Reiser, Rick Szeliski, Dor Verbin, Pratul Srinivasan, Ben Mildenhall, Andreas Geiger, Jon Barron, and Peter Hedman. Merf: Memory-efficient radiance fields for real-time view synthesis in unbounded scenes. *ACM Transactions on Graphics (TOG)*, 42(4):1–12, 2023. 8

[83] Barbara Roessle, Jonathan T. Barron, Ben Mildenhall, Pratul P. Srinivasan, and Matthias Nießner. Dense depth priors for neural radiance fields from sparse input views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 16

[84] Erik Sandström, Yue Li, Luc Van Gool, and Martin R. Oswald. Point-slam: Dense neural point cloud-based slam. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2, 5, 6, 16

[85] Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From coarse to fine: Robust hierarchical localization at large scale. In *CVPR*, 2019. 6, 18

[86] Carolin Schmitt, Simon Donne, Gernot Riegler, Vladlen Koltun, and Andreas Geiger. On joint estimation of pose, geometry and svbrdf from a handheld scanner. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[87] Carolin Schmitt, Božidar Antić, Andrei Neculai, Joo Ho Lee, and Andreas Geiger. Towards scalable multi-view reconstruction of geometry and materials. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2

[88] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016. 18

[89] Aarrushi Shandilya, Benjamin Attal, Christian Richardt, James Tompkin, and Matthew O'toole. Neural fields for structured lighting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3512–3522, 2023. 2

[90] Boxin Shi, Zhe Wu, Zhipeng Mo, Dinglong Duan, Sai-Kit Yeung, and Ping Tan. A benchmark dataset and evaluation for non-lambertian and uncalibrated photometric stereo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3707–3716, 2016. 5

[91] Shining3D. Einscansp, 2023. 2

[92] Sketchfab. Sketchfab: Online 3d geometry viewer, 2024. 17, 18

[93] Pratul P. Srinivasan, Stephan J. Garbin, Dor Verbin, Jonathan T. Barron, and Ben Mildenhall. Nuvo: Neural uv mapping for unruly 3d representations. *arXiv*, 2023. 8, 17

[94] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, Anton Clarkson, Mingfei Yan, Brian Budge, Yajie Yan, Xiaqing Pan, June Yon, Yuyang Zou, Kimberly Leon, Nigel Carter, Jesus Briales, Tyler Gillingham, Elias Mueggler, Luis Pesqueira, Manolis Savva, Dhruv Batra, Hauke M. Strasdat, Renzo De Nardi, Michael Goesele, Steven Lovegrove, and Richard Newcombe. The Replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 6

[95] Jiaming Sun, Xi Chen, Qianqian Wang, Zhengqi Li, Hadar Averbuch-Elor, Xiaowei Zhou, and Noah Snavely. Neural 3d reconstruction in the wild. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–9, 2022. 2, 14

[96] Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in Neural Information Processing Systems*, 33:7537–7547, 2020. 2, 13

[97] Matthew Tancik, Ethan Weber, Evonne Ng, Ruilong Li, Brent Yi, Justin Kerr, Terrance Wang, Alexander Kristoffersen, Jake Austin, Kamyar Salahi, Abhik Ahuja, David McAllister, and Angjoo Kanazawa. Nerfstudio: A modular framework for neural radiance field development. In *ACM SIGGRAPH 2023 Conference Proceedings*, 2023. 1, 15, 17

[98] Marco Toschi, Riccardo De Matteo, Riccardo Spezialetti, Daniele De Gregorio, Luigi Di Stefano, and Samuele Salti. Relight my nerf: A dataset for novel view synthesis and relighting of real world objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20762–20772, 2023. 5, 8, 16

[99] Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 13(04): 376–380, 1991. 18

[100] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021. 2, 3, 8, 13, 14, 15, 17

[101] Zian Wang, Tianchang Shen, Merlin Nimier-David, Nicholas Sharp, Jun Gao, Alexander Keller, Sanja Fidler, Thomas Müller, and Zan Gojcic. Adaptive shells for efficient neural radiance field rendering. *ACM Trans. Graph.*, 42(6), 2023. 2, 4, 8, 14, 15

[102] Sven Woop, Louis Feng, Ingo Wald, and Carsten Benthin. Embree ray tracing kernels for cpus and the xeon phi architecture. In *ACM SIGGRAPH 2013 Talks*, pages 1–1. 2013. 17

[103] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, and Dacheng Tao. Gmflow: Learning optical flow via global matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8121–8130, 2022. 2, 3, 5, 14, 19

[104] Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. *Computer Vision and Pattern Recognition (CVPR)*, 2020. 5

[105] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. *Advances in Neural Information Processing Systems*, 34:4805–4815, 2021. 2, 3, 4, 6, 13, 14

[106] Lior Yariv, Peter Hedman, Christian Reiser, Dor Verbin, Pratul P Srinivasan, Richard Szeliski, Jonathan T Barron, and Ben Mildenhall. Bakedsdf: Meshing neural sdfs for real-time view synthesis. *arXiv preprint arXiv:2302.14859*, 2023. 4

[107] Jonathan Young. Sketchfab: Online 3d geometry viewer, 2024. 8, 17

[108] Zehao Yu, Songyou Peng, Michael Niemeyer, Torsten Sattler, and Andreas Geiger. Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 1, 2, 3, 13, 15

[109] Ramin Zabih and John Woodfill. Non-parametric local transforms for computing visual correspondence. In *Computer Vision—ECCV'94: Third European Conference on Computer Vision Stockholm, Sweden, May 2–6 1994 Proceedings, Volume II 3*, pages 151–158. Springer, 1994. 6, 19

[110] ZDNet, 2023. 2

[111] Kai Zhang, Fujun Luan, Zhengqi Li, and Noah Snavely. Iron: Inverse rendering by optimizing neural sdfs and materials from photometric images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5565–5574, 2022. 2, 7

[112] Yinda Zhang, Neal Wadhwa, Sergio Orts-Escolano, Christian Häne, Sean Fanello, and Rahul Garg. Du 2 net: Learning depth estimation from dual-cameras and dual-pixels. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 582–598. Springer, 2020. 2

[113] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Fast global registration. In *European Conference on Computer Vision*, pages 766–782. Springer, 2016. 6, 18

[114] Zhenglong Zhou, Zhe Wu, and Ping Tan. Multi-view photometric stereo with spatially varying isotropic materials. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1482–1489, 2013. 2

[115] Michael Zollhöfer, Angela Dai, Matthias Innmann, Chenglei Wu, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Shading-based refinement on volumetric signed distance functions. *ACM Transactions on Graphics (ToG)*, 34(4):1–14, 2015. 3