The Sterkfontein Caves Dataset: A Novel View Synthesis Challenge from the Cradle of Humankind

Ireton Liu[1], Brian Xu[2], Steven James[1,3], Dominic Stratford[1], Richard Klein[1,3], and James Tompkin[2]

[1] School of Computer Science and Applied Mathematics, University of the Witwatersrand, Johannesburg, South Africa
[2] Brown University, Providence, RI, USA
[3] Machine Intelligence and Neural Discovery (MIND) Institute, University of the Witwatersrand, Johannesburg, South Africa.

Abstract. We introduce the challenging *Sterkfontein Caves* dataset comprising ten underground scenes from a UNESCO World Heritage Site, and use it to find a new simple baseline method that beats existing low-light reconstruction methods upon it. Each scene is of complex surface geometry and high-frequency texture from cave rock structures, including human-made markings on the rock face. The captured images exhibit varied or uncontrolled lighting over a large dynamic range, with glare artefacts, and low signal-to-noise ratios from the challenging dark real-world capture scenario. We propose a view synthesis benchmark for low-light RAW and sRGB reconstruction. All tested NeRF and Gaussian splatting baseline methods struggle on this data, with the best performing method in terms of reliability and average PSNR being our Raw-Nerfacto method. We discuss these errors in detail to find directions of future work for the community in overcoming the significant challenges that remain in low-light high-detail scenes. Our dataset is publicly available at https://visual.cs.brown.edu/sterkfontein.

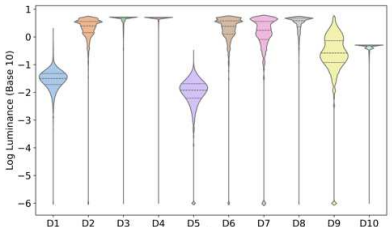
1 Introduction

A long-term goal of view synthesis is to enable global virtual access to culturally important, remote, sensitive, or inaccessible environments. The Sterkfontein Caves are one such environment: a UNESCO World Heritage Site and South African National Heritage Site that is one of the world's richest palaeoanthropological sites.

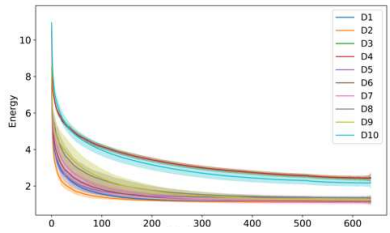
Excavation in the ancient caves has yielded the largest collection of Australopithecus fossils, including iconic specimens like Sts 5 ("Mrs. Ples"), and the most complete early hominin skeleton, StW 573 ("Little Foot") [27]. Beyond the rich record of stone tools, hominin and non-hominin fauna, and fossil wood, the caves represent important karst topographies containing diverse and complex geological formations such as speleothems. Visual documentation of this environment is valuable across several domains: it facilitates science communication [26], provides virtual tourism, enables heritage and ecological conservation monitoring, and aids primary palaeoanthropological research as part of exploration and excavation

Table 1: Characteristics and parameters of the ten scenes. Lighting: the dominant type of light source. **Self-Occ.:** if the scene exhibits notable self-occlusion on the surface. **Co-Loc.:** whether the images were taken with a co-located light source. **ISO:** ISO level. **Shutter:** shutter speed measured in seconds. **Views:** number of unique views in the scene.

Location	Label	Characteristics	Lighting	Self-Occ.	Co-Loc.	ISO	Shutter	Views
Deposit1	D1	Glare	Artificial	✓	✗	2000	1/10	99
Deposit2	D2	Contrast Loss	Artificial	✗	✓	2000	1/100	105
Deposit3	D3	Self Occlusion	Artificial	✓	✗	2000	1/100	89
Deposit4	D4	Self Occlusion	Artificial	✓	✓	2000	1/100	87
Tunnel1	D5	Low Light	Natural	✗	✗	3200	1/10	80
Tunnel2	D6	Low Light	Both	✗	✗	2000	1/100	102
Protrusion1	D7	High Dynamic Range	Natural	✓	✗	2000	1/100	175
Protrusion2	D8	High Dynamic Range	Both	✓	✗	2000	1/100	98
RoofPendant1	D9	Low Light; Scale	Artificial	✗	✓	2000	1/10	116
RoofPendant2	D10	Scale	Artificial	✗	✓	2000	1/100	124



(a) Relative log-scale luminance distribution normalised to exposure and ISO per dataset.



(b) Radial frequency profile of scenes shows still high energy at high frequencies.

Fig. 1: Luminance distribution and radial frequency profiles for the ten scenes.

programmes. Novel view synthesis (NVS) is a promising approach to further these efforts by providing an intuitive 3D viewing experience from captured photographs [13, 18, 21].

However, while some NVS methods have addressed low-light conditions [4, 19, 37], the Sterkfontein Caves present a formidable challenge that existing approaches do not adequately handle. Illumination in the caves may come from narrow shafts of natural light or weak artificial sources, creating large unlit or shadowed regions, strong self-shadows, high dynamic range, glare, and low signal-to-noise ratios. Surface geometry and material vary abruptly, producing high-frequency appearance changes. To complicate matters, field geologists and palaeoarchaeologists may only have simple imaging equipment like smartphones and may need to work quickly with limited power and site access. Thus, beyond providing examples of digital preservation for an important global archaeological site, the Sterkfontein Caves dataset also provides a challenging test-bed to evaluate the efficacy of NVS methods and to further their research.

We contribute a dataset and an analysis of NVS methods that reveals a new state-of-the-art baseline. The *Sterkfontein Caves* dataset is the first public multi-view RAW capture of scenes from a UNESCO World Heritage cave system.

We capture ten scenes with varied lighting configurations and camera parameters to create different reconstruction challenges, and include 80-exposure stacks from which we derive a low-noise reference view. Characteristics and details of these scenes are provided by Tab. 1 and Fig. 1.

We evaluate ten methods for RAW and sRGB inputs on this data. These include standard NeRF and GS methods, those explicitly designed for low-light settings, and methods that denoise for low SNR settings. Overall, we find the average performance to be poor, that denoising via learned priors may not generalise to the geological setting and so is often worse than no denoising, and that low light methods are *not* necessarily better than standard methods. Through this, we uncover a simple adapted baseline—Raw-Nerfacto—that consistently produces good results without catastrophic failures and is able to expose scratched writing on rock faces. Importantly, quantitative metrics often mislead in describing errors within low-light imagery; we instead lean on qualitative analysis to identify research directions for the community. Overall, the digital preservation of heritage sites with harsh illumination remains a challenge, and our contributions aim to further research in recording humanity's most important archaeological locations.

2 Related Work

Datasets for Novel View Synthesis. Early benchmarks for neural rendering focused on controlled conditions. Synthetic datasets like Blender scenes [21] and the LLFF collection [20] provide clean imagery with known camera poses, while indoor datasets such as Replica [28] and ScanNet [5] capture structured environments under consistent artificial lighting. Knapitsch et al. [14] introduce outdoor scenes with natural lighting but maintain high luminance and SNR with good visibility.

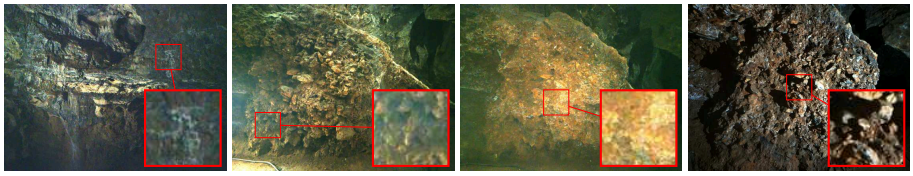
For larger-scale scenes, Tancik et al. [30] demonstrate city-wide reconstruction from internet photos, while MegaScenes [33] provides diverse outdoor imagery. The OMMO dataset [17] combines RGB and point clouds for multi-modal reconstruction, and recent benchmarks like GigaNVS [34] target unbounded, high-resolution capture. Despite their scale, these datasets again feature well-illuminated scenes with good visibility conditions.

Heritage and archaeological site documentation remains underexplored in the neural rendering literature. Existing methods rely on photogrammetry, terrestrial laser scanning [7, 10], and lidar-based cave mapping techniques [6, 42], which prioritise geometric accuracy over photorealistic appearance modelling. These include complex robotics mapping problems in demanding circumstances [38, 39] and part of the DARPA Subterranean Challenge [32]. But, to our knowledge, no dataset exists that benchmarks NVS methods in challenging lighting conditions while providing scientific relevance in adjacent fields such as palaeoarchaeology.

Low-Light Neural Rendering. Standard NeRF and 3DGS pipelines suffer degradation under low-light conditions due to nonlinear tone mapping that introduces multiview inconsistencies and noise amplification in shadows. RawNeRF [19] addresses these limitations by training directly on linear RAW sensor data, preserving high dynamic range while leveraging multi-view aggregation

(a) Low SNR frame **(b)** High SNR HDR+ **(c)** Gaussian-DK **(d)** Luminance-GS

Fig. 2: RAW Sterkfontein data are low SNR. a) One of our RAW frames and b) its HDR+ reference created from 100 exposures, with both post-processed to sRGB. c) and d) are training images from Gaussian-DK [37] and Luminance-GS LOM [3] scenes provided as sRGB (no RAW) using black box camera ISP processing. This limits research opportunities into better noise modelling.



(a) High-frequency **(b)** Glare **(c)** Contrast loss **(d)** Self occlusion

Fig. 3: Characteristic challenges of natural cave datasets. (a) High-frequency surface detail in low signal to noise ratios; (b) glare from static light sources in low-light environments; (c) contrast loss (no shading or shadows) due to co-located light with camera; (d) self occlusion from highly variable and irregular surface geometry. sRGB processed images from our dataset.

to effectively denoise imagery. Li et al. [15] introduce an in-the-loop denoising scheme for Gaussians to also support RAW imagery, and LE3D [11] adds depth regularizations and also supports RAW imagery. LITA-GS [41] employs structural priors and a lightweight progressive denoising model. LL-Gaussian [29] decomposes the scene into static and dynamic components to account for illumination effects. Ye et al. [37] separate rendering into radiance and light feature maps for fine-grained brightness adjustment. DarkGS [40] accounts for inconsistent lighting by calibrating an in-the-loop illumination model. Alternative approaches enhance illumination within the neural rendering pipeline. Aleth-NeRF [4] models light attenuation explicitly in the radiance field, and LLNeRF [35] applies illumination enhancement during reconstruction. For Gaussian splatting, Luminance-GS [3] introduces per-view color mapping and adaptive tone curves for low-light conditions. Still, as we shall see, many methods struggle in underground caves.

3 Sterkfontein Caves Dataset Capture

Acquisition. We pick a cheap, small camera as our capture device. The rationale is that practical use of NVS in caves increases as the capture cost decreases. While larger professional equipment such as 3D scanners or multi-camera rigs with structured light sensors can produce high-quality results, these are more

expensive, more difficult to deploy underground, and out of reach for many field geologists or palaeoarchaeologists. As such, we use a Samsung Galaxy S24 smartphone with Open Camera [8] to capture ten datasets across four different locations. Each image is captured RAW at 4k×3k pixel resolution from the 'Ultrawide' camera, which has a 1/2.55" 12 MP 10-bit sensor. Each scene consists of 80–175 views, so the whole dataset has 1,100+ images. We adjusted camera parameters per scene to balance available light with motion blur: ISO values ranged from 2000–3200 and exposure times ranged from 1/100s to 1/10s. For each viewpoint, following Mildenhall et al. [19], we capture and merge burst sequences of 80–100 frames into a low-noise HDR+ [9] image as a noise-reduced reference (Fig. 2). For each scene, we estimate the camera poses using COLMAP [24] on full-resolution JPEG images postprocessed from our RAW datasets, including a single frame from the burst of 100 for the HDR+ reference.

Scenes. Our ten scenes across four locations represent distinct challenges to NVS (Tab. 1). This is comparable in scale to other NVS benchmarks: MipNeRF-360 (9 scenes), RawNeRF (7), Gaussian-DK (12), and Aleth-NeRF (5). The **Deposit** location shows an excavation containing fallen rocks from a prehistoric opening (a promising location for fossil discoveries). The highly uneven surfaces create self-occlusion that is exacerbated by artificial lighting. The **Tunnel** location has minimal illumination of any kind. The **Protrusion** location features the only cave section illuminated by ceiling openings, creating large dynamic ranges where luminance distributions exhibit distinct peaks at overexposed and occluded regions. Finally, the **Roof Pendant** location is a 360-degree capture of a hanging bedrock in a large open chamber, with human writing scratched into its surface. This scene examines larger-scale reconstruction and fine detail under challenging lighting. Each location has capture variants in exposure time and lighting, including low-light conditions with weak artificial sources that produce glare, co-located lighting to mitigate shadows, and strong spotlight illumination. Across these scenes, we identify six NVS challenges:

1) Low light & high noise. Caves are dark with sparse natural light or even pitch black with only sparse artificial light. This scenario is difficult for patterned colour filter imaging as we must handle low SNR while preserving geometric details (Fig. 3a). Some existing low-light datasets for NVS are sRGB with high SNR; they do not expose RAW images and rely on black box ISPs to denoise (Fig. 2). These data largely ignore noise as a problem which limits research opportunities into better noise modelling.

2) High frequency. Caves have high frequency geometry and texture detail everywhere (Fig. 3a).

3) Self occlusion. Complex cave geometry with uneven surfaces (Fig. 3d) and overhangs creates extensive shadowing, exacerbated by directional artificial lighting. These conditions challenge methods to infer occluded appearance.

4) Contrast loss. Positioning light sources near the camera reduces self-occlusion while improving scene illumination. However, when a fixed light source is near the camera, it causes a loss of contrast as shading and shadows are filled in, reducing information used for geometry estimation (cf. Fig. 3b and Fig. 3c).

5) High dynamic range. Sections of the caves receive natural light from entrances or ceiling openings at certain times of the day. Midday sunlight serves as a strong directional source but illuminates limited sections only, causing the unlit sections of the caves to be significantly darker.
6) Glare. The few static light sources in the caves that provide minimal illumination for traversal cause noticeable glare in images captured nearby (Fig. 3b).

Lighting setups in general are difficult in caves. Dynamic co-located lighting is often unavoidable. Even without dynamic co-located lighting, in a dark sparsely-lit environment, static lighting can introduce prominent shadows cast by the camera operator, and even with great care it is difficult to suppress lighting variation caused by minor occlusions from the operator. These violate the multi-view consistency assumptions underlying many NVS methods, making reconstruction more challenging. Such variations must be accommodated within the reconstruction algorithm.

4 Experimental Setup

We prepare training and test data for RAW methods and methods that input processed images. To generate test hold-out views, we follow the HDR+ algorithm. We average the burst of 80-100 exposures to produce a single RAW denoised image. Then, we demosaic it using bilinear interpolation, transform the data to the CIE XYZ colour space and then to sRGB primaries, and *do not* apply gamma correction such that the image remains linear ('linear sRGB'). Image intensities are divided by 99th percentile of peak luminance and clipped to [0,1].

To generate training views for processed non-RAW methods, we follow the same approach as for the reference view without multi-exposure averaging. Training and evaluation happen on so-called linear sRGB imagery.

Training views for RAW methods simply use the RAW file. Rendered views from RAW reconstruction methods have all camera primary RGB colours per pixel, and reconstruction uses a loss that penalises each original RAW colour filter per pixel. To evaluate outputs, we process the rendered RAW view as above such that it matches the hold-out view.

4.1 Evaluated methods.

RawNeRF [19] RAW input. NeRF method for low light / HDR. This method is MipNeRF [1] with a modified training loss for RAW images.
Nerfacto [31] / Raw-Nerfacto Linear sRGB / RAW input. We adapt the underlying architecture of Nerfacto, itself an adaptation of InstantNGP [22], to take input RAW images and penalise a RAW training loss as in RawNeRF. To our knowledge, this method has not been previously evaluated.
Splatfacto [31] / RawSplatfacto Linear sRGB / RAW input. 3D Gaussian Splatting [13] as implemented in NeRFStudio as Splatfacto, then adapted by us to take as input unprocessed RAW images and optimise against them. Again, the RAW version has not been evaluated to our knowledge.

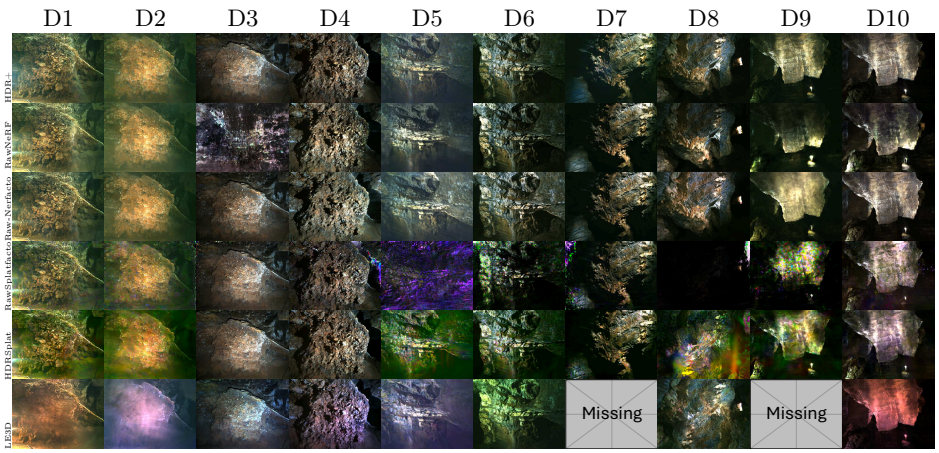


Fig. 4: Qualitative comparison of all RAW methods. Raw-Nerfacto is visibly better than its competitors, which often cannot reliably reconstruct challenging scenes.

LE3D [11] RAW input. 3DGS method for low light with additional random points for initialisation, a colour MLP instead of spherical harmonics, and depth and blending weight regularisation during training.
HDRSplat [25] RAW input. HDRSplat adapts the base 3DGS implementation and also adds a learning-based RAW image denoiser PMRID [36] with hand-tuned rasterisation parameters for training on linear RAW images.
Gaussian-DK [37] Linear sRGB input. 3DGS method for low light that splits the GS rasterisation process into radiance and light field maps, which are separately tone-mapped to account for brightness variations in images.

We also include two low-light enhancement methods:
Aleth-NeRF [4] Linear sRGB input. NeRF with an illumination-adaptive "concealing" field that causes scene darkness to obstruct light to the camera.
Luminance-GS [3] Linear sRGB input. This augments 3D Gaussian Splatting with per-view colour matrix mapping and view-adaptive curve adjustment to account for challenging lighting conditions.

4.2 Evaluation protocol

For each method, we conduct experiments using the official code release from the authors with default hyperparameters. Each method is trained and evaluated on full-resolution 4k×3k images, rather than on the often-downsampled imagery in other datasets. Training utilises all available views, which vary in number across scenes (see Table 1). For evaluation, we hold out a single view per scene that shares the exact perspective of the HDR+ reference frame, which is inherently denoised via exposure stacking. During evaluation, we render the held-out view at full sensor resolution and compare it against the HDR+ reference. Comparison is done in the linear SRGB space where possible. NeRF-based methods are trained for 200k iterations; 3DGS-based methods for 30k.

Table 2: Quantitative comparison of all RAW methods using PSNR, SSIM, and LPIPS metrics. Blue = best; yellow = 2nd best.

Method	Metric	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
RawNeRF *RAW in+out*	PSNR▲	18.1	**19.4**	12.1	14.4	**17.2**	18.6	18.1	17.4	13.7	15.1
	SSIM ▲	**0.251**	**0.241**	0.322	0.382	**0.175**	**0.420**	**0.403**	0.573	0.487	0.568
	LPIPS▼	0.740	0.800	0.794	0.567	0.896	0.701	0.619	0.558	0.670	0.593
Raw-Nerfacto *RAW in+out*	PSNR▲	**18.4**	19.2	19.6	14.0	15.6	18.0	**18.8**	**21.3**	16.9	18.9
	SSIM ▲	0.24	0.229	0.510	0.323	0.163	0.412	0.400	**0.621**	**0.504**	0.597
	LPIPS▼	0.699	**0.75**	0.413	0.542	0.891	0.711	**0.605**	**0.519**	**0.616**	**0.459**
Raw-Splatfacto[1] *RAW in+out*	PSNR▲	17.9	17.4	19.6	13.6	12.0[1]	14.4[1]	16.7	11.7	15.3	17.8[1]
	SSIM ▲	0.234	0.182	0.495	0.310	0.109[1]	0.179[1]	0.330	0.113	0.191	0.501[1]
	LPIPS▼	**0.695**	0.782	0.447	0.600	0.897[1]	0.743[1]	0.629	0.701	0.741	0.648[1]
HDRSplat *RAW in+out*	PSNR▲	16.9	17.4	**20.1**	15.2	14.2	18.8	17.5	12.4	**17.5**	18.7
	SSIM ▲	0.218	0.162	**0.526**	0.384	0.111	0.267	0.182	0.289	0.353	0.566
	LPIPS▼	0.775	0.835	**0.379**	**0.464**	0.932	0.676	0.645	0.736	0.73	0.585
LE3D *RAW in+out*	PSNR▲	17.0	13.7	19.3	**15.3**	16.0	**19.0**	–[2]	17.1	–[2]	**19.3**
	SSIM ▲	0.244	0.216	0.517	**0.405**	0.167	0.395	–[2]	0.447	–[2]	0.650
	LPIPS▼	0.859	0.908	0.466	0.650	**0.854**	**0.659**	–	0.650	–	0.593

[1] Scene that completed fewer than 30,000 training steps due to OOM caused by uncontrolled Gaussian growth.
[2] Scene terminated due to out-of-memory errors (OOM) before training could begin.

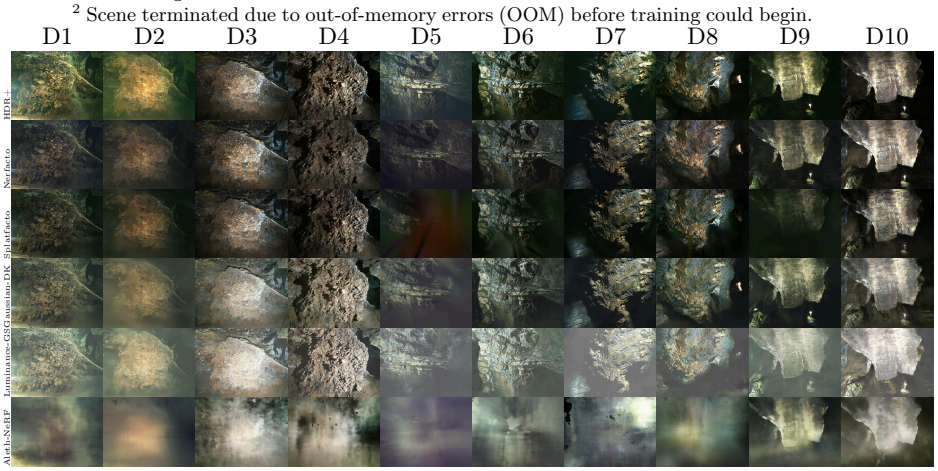


Fig. 5: Qualitative comparison of all sRGB methods.

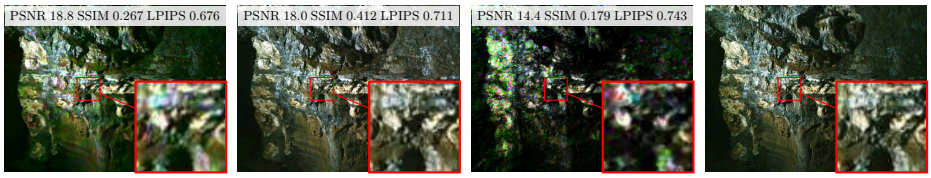
All experiments were conducted on NVIDIA RTX 3090s with 24GB VRAM or NVIDIA Quadro RTX8000 with 48GB VRAM.

We compute PSNR, SSIM, LPIPS metrics. Some methods directly output tone-mapped images (Gaussian-DK, Luminance-GS, Aleth-NeRF) via internal tone mapping. This step can be sophisticated, e.g., a CNN in Gaussian-DK, which makes appearance matching across techniques for visualisation and comparison difficult, and which makes numeric metric comparisons across methods and scenes less meaningful. Further, Luminance-GS and Aleth-NeRF attempt to artificially brighten the scene, e.g., in Sec. 4.1, the output tone of Luminance-GS is bright and varies across scenes, which again makes their quantitative results

Table 3: Comparison of all sRGB methods qualitatively and using PSNR, SSIM, and LPIPS metrics. Blue = best; yellow = 2nd best.

Method	Metric	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
Nerfacto	PSNR▲	14.6	14.6	**19.4**	**17.5**	14.1	**17.5**	**18.3**	**20.1**	**17.8**	16.9
Linear sRGB in	SSIM ▲	0.215	0.202	**0.479**	**0.448**	0.144	**0.380**	**0.416**	**0.510**	0.439	0.490
Linear sRGB out	LPIPS▼	**0.609**	**0.692**	**0.324**	**0.452**	**0.709**	**0.534**	**0.521**	**0.398**	**0.542**	**0.460**
Splatfacto	PSNR▲	13.8	14.0	19.1	14.9	12.2	15.7	17.2	18.9	13.6	**19.1**
Linear sRGB in	SSIM ▲	0.210	0.198	0.488	0.365	0.127	0.350	0.387	0.499	**0.476**	**0.504**
Linear sRGB out	LPIPS▼	0.716	0.815	0.366	0.471	0.904	0.625	0.585	0.537	0.667	0.467
Learned CNN tone mapper may produce different results from reference tone mapper.											
Gaussian-DK	PSNR▲	**17.7**	**17.7**	13.3	12.2	**16.6**	16.1	16.9	17.1	16.9	15.2
Linear sRGB in	SSIM ▲	**0.239**	**0.215**	0.333	0.227	**0.162**	0.287	0.380	0.408	0.427	0.382
Tone map out	LPIPS▼	0.755	0.808	0.517	0.542	0.836	0.716	0.650	0.581	0.701	0.612
Low light enhancement methods[2]											
Luminance-GS	PSNR▲	13.3	13.7	9.6	10.1	12.6	10.2	7.9	10.6	9.6	8.7
Linear sRGB in	SSIM ▲	0.200	0.188	0.262	0.182	0.131	0.167	0.152	0.229	0.202	0.218
Brightened out	LPIPS▼	0.722	0.783	0.470	0.539	0.818	0.709	0.744	0.677	0.772	0.663
Aleth-NeRF[1]	PSNR▲	15.9	16.0	10.7	9.6	14.0	10.6	10.0	12.2	13.5	12.6
Linear sRGB in	SSIM ▲	0.225	0.211	0.373	0.210	0.143	0.185	0.225	0.309	0.298	0.324
Brightened out	LPIPS▼	0.779	0.824	0.751	0.820	0.864	0.838	0.837	0.762	0.682	0.727

[1] Aleth-NeRF did not reasonably represent any scene; metrics presented here must be interpreted with caution.
[2] Metrics are not as meaningful as images brightened; included for completeness



(a) HDRSplat **(b)** Raw-Nerfacto **(c)** Raw-Splatfacto **(d)** HDR+ Ref

Fig. 6: Standard metrics do not measure artefacts well. HDRSplat's render (a) is noisier throughout the image than that of Raw-Nerfacto (b), yet (a)'s PSNR is better. LPIPS varies only moderately with large visual changes, where HDRSplat is incorrectly better than Raw-Nerfacto while Splatfacto is correctly worse. SSIM better reflects the difference across (a–c).

less meaningful. We include them for completeness, nonetheless, and direct the reader to the qualitative results (Fig. 11).

5 Results and Discussion

First, we remind ourselves that the quantitative assessment of images is difficult. Figure 6 shows that existing standard metrics alone are insufficient for quality assessment, particularly when the reconstruction may have many local errors caused by a low Signal-to-Noise Ratio (SNR). This is particularly true for 3DGS-based methods, as we expect: overall 3DGS-based methods perform worse than NeRF-based methods given noisy input as it is easier for primitives

to overfit. This is to say nothing of its rendering speed; while InstantNGP can render interactively, 3DGS' primary attribute is its faster rendering.

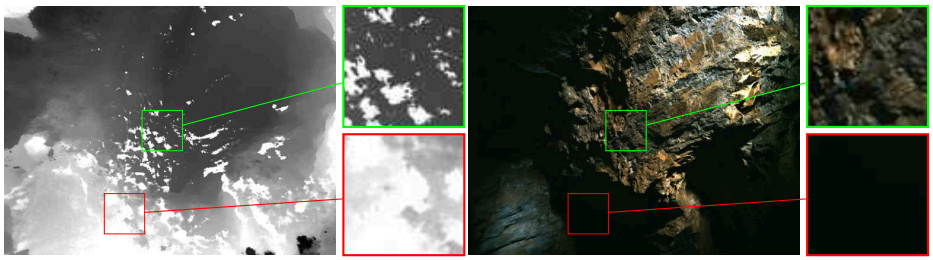
Next, a failure case: Aleth-NeRF does not reconstruct any scene such that renderings are coherent, with only a coarse depiction for D9 and D10. It is built upon the original NeRF architecture (not MipNeRF-360), and the optimisation fails to converge on our data.

RAW methods. Figure 4 and Tab. 2 contain the quantitative results of all the RAW methods benchmarked. Even though it is a simple adaptation, Raw-Nerfacto is the most effective at reconstructing our challenging datasets, often performing better than RawNeRF and without catastrophic failures (e.g., RawNeRF on D3). It is also an order of magnitude faster to train and infer on average, which can be attributed to the sampling efficiency of the multi-resolution hash table [22] and TinyCudaNN [23]. Raw-Splatfacto performs reasonably on some scenes (D1, D4, D10) but exhibits catastrophic failure in very low light (D5) or with large dynamic range (D8; peak luminance is significantly overestimated which causes the qualitative image to look dark). Further, results have many minor artefacts with bright colours (e.g., purple) fit across clusters of high noise RAW pixels.

LE3D overall also struggles on this data. While it performs better on higher luminance scenes (D3) in a similar way to other methods, and while it does produce correct large-scale features otherwise, it incorrectly fits primitives to basic surface details that other GS methods like Raw-Splatfacto do not struggle with. Some scenes (D7, D9) fail to even start due to high memory requirements: LE3D replaces spherical harmonics with a per-Gaussian MLP, which significantly increases per-primitive memory, and high image frequency further exacerbates memory use in 3DGS [16]. We choose not to downsample images to save memory, such that we maintain a like-for-like comparison in the RAW image domain.

The "denoise first" approach of HDRSplat using PMRID [36] was ineffective on our low-luminance scenes (Fig. 9). This is likely because our datasets are out-of-distribution for the PMRID model, which was trained on the SID dataset [2], and comprises mostly urban imagery (household, street scenes); more classical approaches may fare better [12]. Comparing scenes with different SNR due to varying ISO levels (D1 and D5), there is little difference in PMRID performance, suggesting the failure is due to domain shift rather than noise level alone. On scenes with more light, HDRSplat appears quantitatively competitive with other RAW processing methods (D3, D4) and qualitatively also on D6. On these scenes, the success demonstrates the potential benefit of denoising afforded by HDRSplat when applied to 3DGS. However, this is brittle: other scenes show significant artefacting (Fig. 6) that overall reduces quality unacceptably (D2, D5, D8, D9).

RAW vs. linear sRGB methods. With equivalent methods, comparing RAW to sRGB isolates the impact of the imaging pipeline on reconstruction. For example, as shown in Tab. 2 and Tab. 3, Raw-Nerfacto is on average better than Nerfacto numerically (18.07 vs. 17.08 dB). However, as GS suffers under noise, Raw-Splatfacto is numerically worse than Splatfacto (14.36 vs. 15.85 dB). Qualitatively, however, if luminance is high enough, Raw-Splatfacto can recover detail despite the noise (D2) through multiview constraints, whereas demosaiced Splat-



(a) Depth Est. **(b)** sRGB prediction

Fig. 7: Self-shadowing effects on D7 (cropped). a) Self-occluded regions exhibit missing depth density, indicating an issue in geometric reconstruction. b) Raw-Nerfacto's rendering produces visually coherent results.

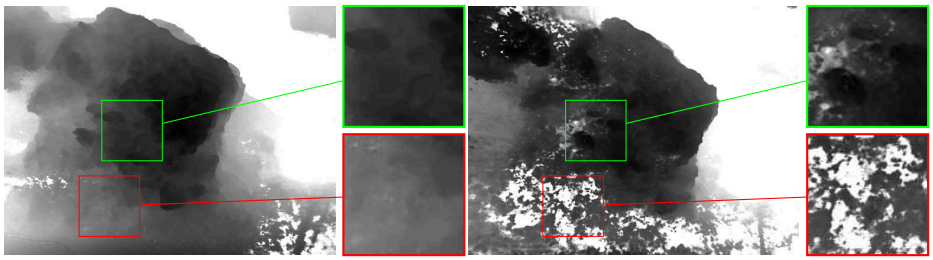
facto often yields blurred results in these areas and in darker regions. Nonetheless, RAW introduces more scene-level failures, with other regions showing poorer detail recovery (D9).

Challenge: Low light & high noise. While Gaussian-DK, Aleth-NeRF, and Luminance-GS all show improved low-light scene prediction in their respective studies, their data do not exhibit high noise. This causes large, blurry regions in their reconstructions on our data, as well as ill-fitted high-frequency Gaussians that are not able to capture the underlying scene texture (Sec. 4.1). Given the less meaningful quantitative metrics for these methods, careful qualitative inspection shows that Gaussian-DK and Luminance-GS can perform slightly better than Splatfacto (D5, D6) though often produce comparable results (all others).

Challenge: High frequency. The naturally high frequency appearance of the caves is not well captured by most methods. With the exception of Raw-Nerfacto, which qualitatively produced the sharpest results most often, other renderings that look visibly "softer" than the reference images, and even Raw-Nerfacto fails to achieve high PSNR. In regions where 3DGS methods appear to have captured the high frequency details, closer inspection reveals tell-tale needle like artefacts from mal-fit Gaussian primitives.

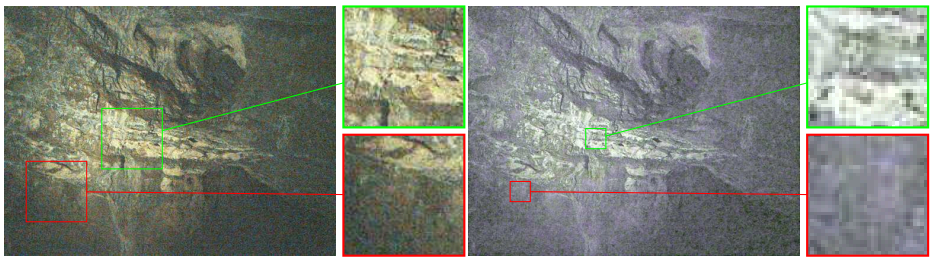
Challenge: Self occlusion. Surprisingly, reconstruction of shadowed and occluded areas (present in D3 and D4) only became a major hurdle in D7, where the dynamic range and extent of occlusion were much greater. NeRF-based methods like Raw-Nerfacto failed to estimate density in these dark regions (Fig. 7), and 3DGS-based methods similarly struggle or fit to noise. As shadow interiors provide little signal, they may require careful completion instead.

Challenge: Contrast loss. This is caused by any light source oriented to illuminate the camera's viewpoint. This fills in shadows and produces uniform shading, reducing image contrast and, in turn, adversely impacting depth estimation. When examining the depth maps (D1 and D2) rendered by Raw-Nerfacto (Fig. 8), this failure manifests as low-density "holes" on the surface of the protrusion, alongside poor shape estimation of the surrounding geometry. Other methods exhibit failure modes as flat, blurry surfaces in the central region of the scene, where methods would typically produce the most detail. Consequently, a



(a) Depth estimate of D1 **(b)** Depth estimate of D2

Fig. 8: Contrast loss causes floaters and gaps. Depth estimation of a) D1 and b) D2 using Raw-Nerfacto. D2 has reduced contrast due to a light appearing near the camera. This introduces floaters around the center and low-density areas.



(a) Noisy training image. **(b)** Result of PMRID.

Fig. 9: Learned denoisers need training on caves. Comparison showing (a) a noisy input image and (b) the result of the PMRID method applied to the RAW data, which creates blobby artefacts and adds purple. This outcome affects reconstruction, highlighting a limitation of this denoising approach.

noticeable degradation in depth reconstruction quality appears in D1 and D2, regardless of the underlying reconstruction technique.

Challenge: High-dynamic range. In D7 and D8, signals from the brightly lit section overwhelm the already faint signals in the occluded areas. Furthermore, the signal-to-noise ratio (SNR) in these occluded regions is particularly low relative to the rest of the scene. Consequently, all methods completely fail to estimate the geometry in these areas coherently. The weak yet visibly present signal becomes difficult for the model to predict; while NeRF-based methods fail to predict any density values (Fig. 7a), 3DGS methods instead place a large, textureless Gaussian.

Challenge: Glare. Ideally, we could remove glare as it is not interesting to geologists or palaeoarchaeologists. As a physical effect, glare depends primarily upon the angle of the camera to the light, and can appear relatively fixed within the visual field as the camera makes small translations. Thus, glare appears in every reconstruction for every method as part of the geometry. In 3DGS-based methods, it is typically represented by large Gaussians placed near the location of the light source. In NeRF-based methods, in addition to adding low-density values near the light source (Fig. 8a), additional floater-like artefacts appear. Future work should consider how to remove glare.

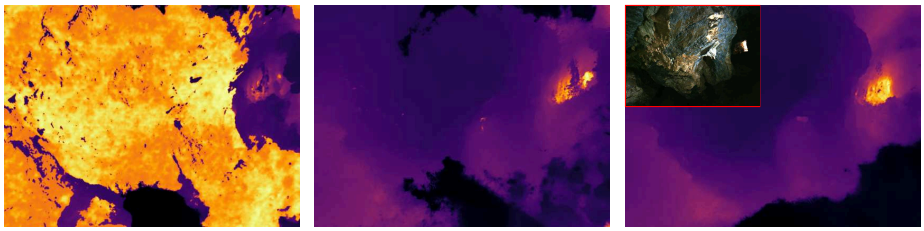
Why might Raw-Nerfacto work better? Raw-Nerfacto is a simple adaptation, yet it outperforms purpose-built low-light methods. We attribute this to a combination of factors that address failure modes in other methods:

1. *RAW supervision preserves a linear noise model.* Demosaicing and ISP processing introduce spatially correlated, non-linear artefacts that break multi-view photometric consistency. Linear RAW data retains approximately Poisson–Gaussian per-pixel noise that is independent across views—good for the multi-view aggregation. On the same architecture, RAW supervision yields ∼1 dB average improvement: Raw-Nerfacto 18.07 vs. Nerfacto 17.08 dB.
2. *Hash encoding as shared regulariser that is robust to low SNR.* Methods differ in how they aggregate information across views during optimisation. Raw-Nerfacto's multi-resolution hash tables [22] share weights across all rays, which resists fitting per-pixel noise while still resolving good spatial detail at multiple scales (Fig. 11). RawNeRF builds on MipNeRF [1], whose integrated positional encoding averages over cone footprints—this may limit high-frequency detail recovery of cave textures. 3DGS primitives have many local degrees of freedom that fit to noise, particularly when adaptive density control spawns small Gaussians (Fig. 4). As discussed above, this is why 3DGS methods struggle.
3. *Sigmoid output activation stabilises training.* RawNeRF uses an exponential output activation to convert from log-space radiance, which is numerically less stable under low SNR. Replacing this with a sigmoid helps stabilise training under high noise, avoiding the catastrophic failures seen with RawNeRF (e.g., D3; Figs. 10a and 10b).
4. *Per-image appearance embeddings (AE).* Raw-Nerfacto uses NeRF-W [18] style per-image appearance embeddings. These can absorb residual variation and prevent it from corrupting the shared 3D representation, e.g., slight lighting occlusions by the camera operator moving within a sparsely-illuminated scene and larger inconsistency caused by dynamic co-located lighting. While other methods may, like Raw-Nerfacto, have per-point features, or may use a learned tone-mapper (Gaussian-DK) or colour matrices (Luminance-GS) on the output, no other compared method affords this per-image flexibility during reconstruction (Fig. 10c).

These points help to explain why a simple adaptation reliably outperforms more complex alternatives on our data.

6 Conclusion

We provide a new field-captured dataset from the Sterkfontein Caves. These data capture high-frequency surface detail among realistic sensor noise, while exposure stacks provide low-noise references. We evaluate NVS methods in this low-light, low-SNR, high-detail setting. On this data, methods designed for low light are not better than standard methods; RAW NeRF methods produce visually better results than regular NeRF ones, yet for 3DGS the reverse holds, and denoising ahead of time is fraught. In the process, we reveal that the previously-unevaluated Raw-Nerfacto method *reliably produces the highest-quality results.* While still low



(a) Exponential **(b)** Sigmoid **(c)** Sigmoid + AE

Fig. 10: Raw-Nerfacto depth maps on D8. (a) Exponential output activation causes severe geometry errors under low SNR. (b) Replacing it with a sigmoid stabilises training. (c) Adding per-image appearance embeddings (AE) further reduces residual artefacts.

in quality overall, they can show detailed human markings on a roof pendant (D10). As the cradle of humankind, the caves themselves are important to document in high-quality 3D; yet no tested method does. Thus, our dataset stands as a challenge to the community to venture underground and solve NVS in the dark.

Limitations. While RAW, our data lacks bracketed exposures for high-dynamic range scenes (D7 and D8). Further, while such equipment is not simple to use by field scientists, geometric ground truth from laser or structured light scanning and material ground truth from gonioreflectometry would expand the set of methods to evaluate NVS. Finally, with more data, ML methods could be promising, though hallucinations are a problem within scientific applications.

Acknowledgements

DS acknowledges the long-term and ongoing support of the Sterkfontein excavations and research from the University of the Witwatersrand, the National Research Foundation (NRF) African Origins Platform (current award number: AOP240512218352), GENUS: DSTI-NRF Centre of Excellence in Palaeosciences, and the Palaeontological Scientific Trust (PAST). We thank the High Performance Computing infrastructure from the Mathematical Sciences Support Unit at the University of the Witwatersrand and the Center for Computation and Visualisation at Brown University for compute. IL thanks the DSI-NRF for financial support (Award Number: PMDS22070735121). BX and JT thank NSF CAREER 2144956. This work was partially supported by the National Geographic Society under Grant NGS-105583T-24.

References

1. Barron, J.T., Mildenhall, B., Tancik, M., Hedman, P., Martin-Brualla, R., Srinivasan, P.P.: Mip-NeRF 360: Unbounded anti-aliased neural radiance fields. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5855–5864 (2021)
2. Chen, C., Chen, Q., Xu, J., Koltun, V.: Learning to see in the dark. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3291–3300 (2018)

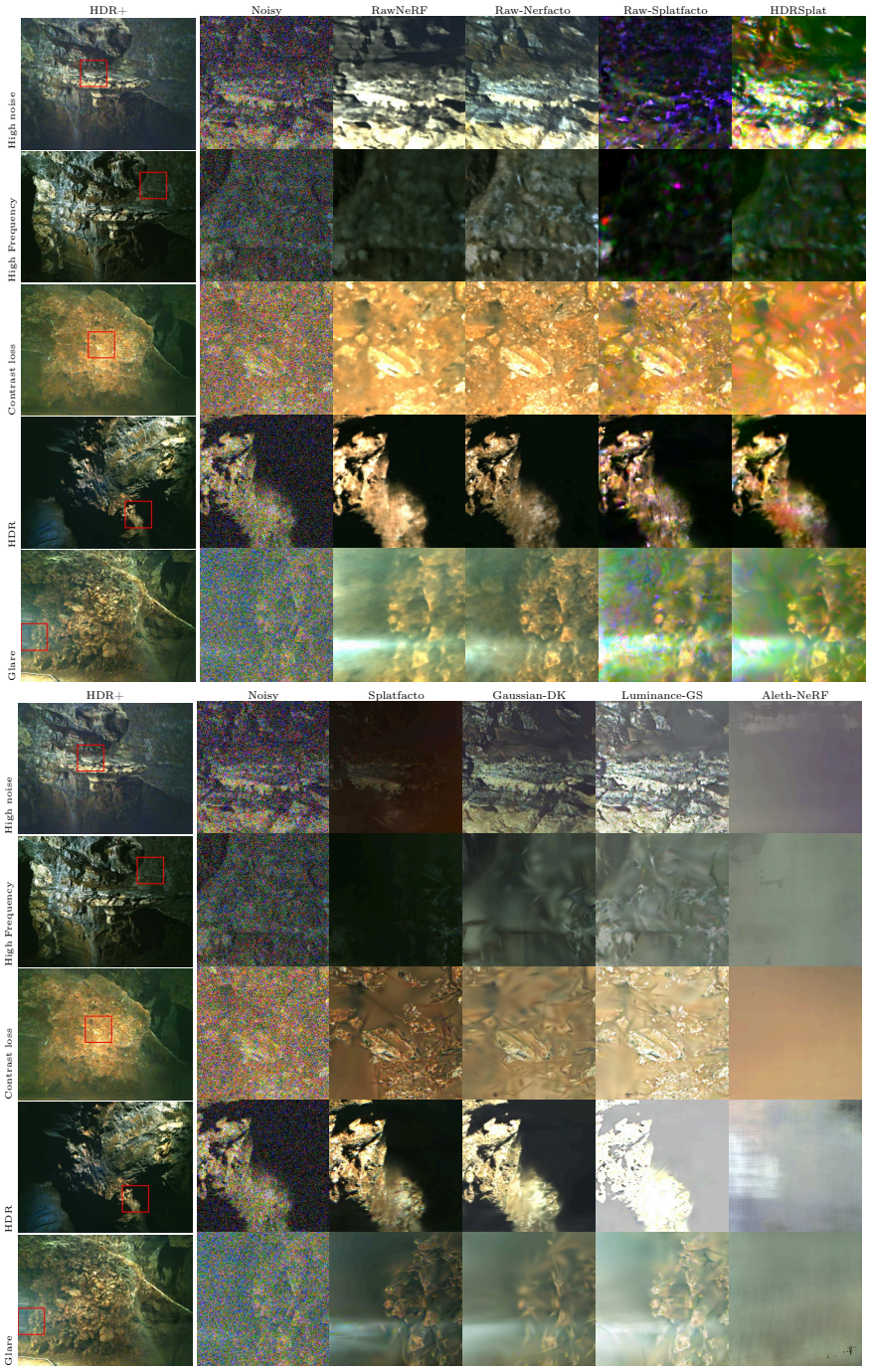


Fig. 11: Reconstruction of challenging areas. From top to bottom: low light & high noise (D5), high frequency geometry and texture (D6), contrast loss (D2), high dynamic range (D7), and glare (D1). Self-occlusion is shown in Fig. 7.

3. Cui, Z., Chu, X., Harada, T.: Luminance-GS: Adapting 3D Gaussian splatting to challenging lighting conditions with view-adaptive curve adjustment. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 26472–26482 (2025)
4. Cui, Z., Gu, L., Sun, X., Ma, X., Qiao, Y., Harada, T.: Aleth-NeRF: Illumination adaptive NeRF with concealing field assumption. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 1435–1444 (2024)
5. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5828–5839 (2017)
6. Fabbri, S., Sauro, F., Santagata, T., Rossi, G., De Waele, J.: High-resolution 3-D mapping using terrestrial laser scanning as a tool for geomorphological and speleogenetical studies in caves: An example from the Lessini mountains (North Italy). Geomorphology **280**, 16–29 (2017)
7. Gallay, M., Kaňuk, J., Hochmuth, Z., Meneely, J.D., Hofierka, J., Sedlák, V.: Large-scale and high-resolution 3-D cave mapping by terrestrial laser scanning: a case study of the Domica Cave, Slovakia. International Journal of Speleology **44**(3), 277–291 (2015)
8. Harman, M.: Open Camera (2024), `sourceforge.net/p/opencamera/code`
9. Hasinoff, S.W., Sharlet, D., Geiss, R., Adams, A., Barron, J.T., Kainz, F., Chen, J., Levoy, M.: Burst photography for high dynamic range and low-light imaging on mobile cameras. ACM Transactions on Graphics **35**(6), 1–12 (2016)
10. Ioannides, M.: Digital Cultural Heritage: Final Conference of the Marie Skłodowska-Curie Initial Training Network for Digital Cultural Heritage, ITN-DCH 2017, Olimje, Slovenia, May 23–25, 2017, Revised Selected Papers, vol. 10605. Springer (2018)
11. Jin, X., Jiao, P., Duan, Z.P., Yang, X., Li, C.Y., Guo, C.L., Ren, B.: Lighting every darkness with 3DGS: Fast training and real-time rendering for HDR view synthesis. In: Proceedings of the 38th International Conference on Neural Information Processing Systems (2024)
12. Jin, X., Niklaus, S., Zhang, Z., Xia, Z., Guo, C., Yang, Y., Chen, J., Li, C.Y.: Classic video denoising in a machine learning world: Robust, fast, and controllable (2025)
13. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3D Gaussian splatting for real-time radiance field rendering. ACM Transaction on Graphics **42**(4), 139–1 (2023)
14. Knapitsch, A., Park, J., Zhou, Q.Y., Koltun, V.: Tanks and temples: Benchmarking large-scale scene reconstruction. ACM Transactions on Graphics **36**(4), 1–13 (2017)
15. Li, Z., Wang, Y., Kot, A., Wen, B.: From chaos to clarity: 3DGS in the dark. Proceedings of the International Conference on Neural Information Processing Systems (2024)
16. Liang, Y., Okunev, M., Uy, M.A., Li, R., Guibas, L., Tompkin, J., Harley, A.W.: Monocular dynamic gaussian splatting: Fast, brittle, and scene complexity rules. arXiv preprint arXiv:2412.04457 (2024)
17. Lu, C., Yin, F., Chen, X., Liu, W., Chen, T., Yu, G., Fan, J.: A large-scale outdoor multi-modal dataset and benchmark for novel view synthesis and implicit scene reconstruction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7557–7567 (2023)
18. Martin-Brualla, R., Radwan, N., Sajjadi, M.S., Barron, J.T., Dosovitskiy, A., Duckworth, D.: NeRF in the wild: Neural radiance fields for unconstrained photo collections. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7210–7219 (2021)
19. Mildenhall, B., Hedman, P., Martin-Brualla, R., Srinivasan, P.P., Barron, J.T.: 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. pp. 16190–16199 (2022)

20. Mildenhall, B., Srinivasan, P.P., Ortiz-Cayon, R., Kalantari, N.K., Ramamoorthi, R., Ng, R., Kar, A.: Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. ACM Transactions on Graphics **38**(4), 1–14 (2019)
21. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: NeRF: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021)
22. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. ACM Transactions on Graphics **41**(4), 1–15 (2022)
23. Müller, T., Rousselle, F., Novák, J., Keller, A.: Real-time neural radiance caching for path tracing. ACM Transactions on Graphics (2021)
24. Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4104–4113 (2016)
25. Singh, S., Garg, A., Mitra, K.: HDRSplat: Gaussian splatting for high dynamic range 3D scene reconstruction from raw images. BMVC (2024)
26. Smith, M., Stratford, D., Bimber, O., Irurah, D.: Towards a virtual environment for the Sterkfontein Caves, South Africa: developing a georeferenced and optimised digital dataset. Antiquity pp. 1–6 (2025)
27. Stratford, D.J.: The Sterkfontein Caves after eighty years of paleoanthropological research: the journey continues. American Anthropologist **120**(1), 39–54 (2018)
28. Straub, J., Whelan, T., Ma, L., Chen, Y., Wijmans, E., Green, S., Engel, J.J., Mur-Artal, R., Ren, C., Verma, S., et al.: The replica dataset: A digital replica of indoor spaces. arXiv preprint arXiv:1906.05797 (2019)
29. Sun, H., Yu, F., Xu, H., Zhang, T., Zou, C.: Ll-gaussian: Low-light scene reconstruction and enhancement via gaussian splatting for novel view synthesis (2025), https://arxiv.org/abs/2504.10331
30. Tancik, M., Casser, V., Yan, X., Pradhan, S., Mildenhall, B., Srinivasan, P.P., Barron, J.T., Kretzschmar, H.: Block-NeRF: Scalable large scene neural view synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8248–8258 (2022)
31. Tancik, M., Weber, E., Ng, E., Li, R., Yi, B., Kerr, J., Wang, T., Kristoffersen, A., Austin, J., Salahi, K., Ahuja, A., McAllister, D., Kanazawa, A.: NeRFstudio: A modular framework for neural radiance field development. In: ACM SIGGRAPH Conference Proceedings. SIGGRAPH '23 (2023)
32. Tranzatto, M., Miki, T., Dharmadhikari, M., Bernreiter, L., Kulkarni, M., Mascarich, F., Andersson, O., Khattak, S., Hutter, M., Siegwart, R., et al.: Cerberus in the DARPA subterranean challenge. Science Robotics **7**(66), eabp9742 (2022)
33. Tung, J., Chou, G., Cai, R., Yang, G., Zhang, K., Wetzstein, G., Hariharan, B., Snavely, N.: Megascenes: Scene-level view synthesis at scale. In: European conference on computer vision. pp. 197–214. Springer (2024)
34. Wang, G., Zhang, J., Wang, F., Huang, R., Fang, L.: XScale-NVS: Cross-scale novel view synthesis with hash featurized manifold. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21029–21039 (2024)
35. Wang, H., Xu, X., Xu, K., Lau, R.W.: Lighting up nerf via unsupervised decomposition and enhancement. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 12632–12641 (2023)
36. Wang, Y., Huang, H., Xu, Q., Liu, J., Liu, Y., Wang, J.: Practical deep raw image denoising on mobile devices. In: European Conference on Computer Vision. pp. 1–16. Springer (2020)

37. Ye, S., Dong, Z.H., Hu, Y., Wen, Y.H., Liu, Y.J.: Gaussian in the dark: Real-time view synthesis from inconsistent dark images using Gaussian splatting. In: Computer Graphics Forum. vol. 43, p. e15213. Wiley Online Library (2024)
38. Zhang, G., Chen, Y., Moyes, H.: Optimal 3D reconstruction of caves using small unmanned aerial systems and RGB-D cameras. In: 2018 International Conference on Unmanned Aircraft Systems (ICUAS). pp. 410–415. IEEE (2018)
39. Zhang, G., Shang, B., Chen, Y., Moyes, H.: SmartCaveDrone: 3D cave mapping using uavs as robotic co-archaeologists. In: 2017 International Conference on Unmanned Aircraft Systems (ICUAS). pp. 1052–1057. IEEE (2017)
40. Zhang, T., Huang, K., Zhi, W., Johnson-Roberson, M.: Darkgs: Learning neural illumination and 3d gaussians relighting for robotic exploration in the dark (2024), https://arxiv.org/abs/2403.10814
41. Zhou, H., Dong, W., Chen, J.: Lita-gs: Illumination-agnostic novel view synthesis via reference-free 3d gaussian splatting and physical priors. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 21580–21589 (2025)
42. Zlot, R., Bosse, M., Greenop, K., Jarzab, Z., Juckes, E., Roberts, J.: Efficiently capturing large, complex cultural heritage sites with a handheld mobile 3D laser mapping system. Journal of Cultural Heritage **15**(6), 670–678 (2014)