

# Dynamic Mesh-Aware Radiance Fields

Yi-Ling Qiao\* Alexander Gao\* Yiran Xu Yue Feng Jia-Bin Huang Ming C. Lin  
University of Maryland College Park

<https://mesh-aware-rf.github.io>



Figure 1: **Mesh-aware rendering of radiance fields.** We place a cubic mesh with reflective textures and other synthetic mesh objects in the *MipNeRF-360 Garden* [5] scene. Our mesh-aware rendering explicitly computes the rays bouncing inside the mirror room, creating an ‘infinite mirror room’ visual effect.

## Abstract

Embedding polygonal mesh assets within photorealistic Neural Radiance Fields (NeRF) volumes, such that they can be rendered and their dynamics simulated in a physically consistent manner with the NeRF, is under-explored from the system perspective of integrating NeRF into the traditional graphics pipeline. This paper designs a two-way coupling between mesh and NeRF during rendering and simulation. We first review the light transport equations for both mesh and NeRF, then distill them into an efficient algorithm for updating radiance and throughput along a cast ray with an arbitrary number of bounces. To resolve the discrepancy between the linear color space that the path tracer assumes and the sRGB color space that standard NeRF uses, we train NeRF with High Dynamic Range (HDR) images. We also present a strategy to estimate light sources and cast shadows on the NeRF. Finally, we consider how the hybrid surface-volumetric formulation can be efficiently integrated with a high-performance physics simulator that supports cloth, rigid and soft bodies. The full rendering and simulation system can be run on a GPU at interactive rates. We show that a hybrid system approach outperforms alternatives in visual realism for mesh insertion, because it allows realistic light transport from volumetric NeRF me-

dia onto surfaces, which affects the appearance of reflective/refractive surfaces and illumination of diffuse surfaces informed by the dynamic scene.

## 1. Introduction

Creating high-quality 3D environments suitable for photorealistic rendering entails labor-intensive manual work carried out by skilled 3D artists. Neural Radiance Fields (NeRF) [46] provide a convenient way to capture a volumetric representation of a complex, real-world scene, paving the way for high-quality novel view synthesis and interactive photorealistic rendering [49]. These qualities make NeRF exceptionally adept at modeling background environments. On the other hand, existing methods for physically-based simulation and rendering of complex material and lighting effects are primarily based on *surface mesh* representations. Integrating neural field representations with well-established traditional graphics pipelines opens up many possibilities in VR/AR, interactive gaming, virtual tourism, education, training, and computer animation.

Volume rendering [56] has demonstrated its capability to produce visually captivating results for participating media [53]. However, integrating NN-based NeRF into this

\*Equal contribution

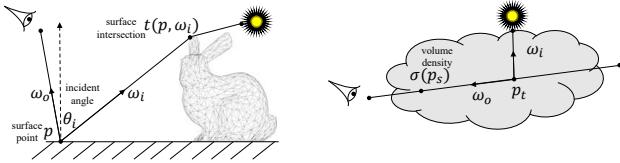


Figure 2: **Light transport** on the surface (*left*) and in the medium (*right*).

pipeline while maintaining realistic lighting effects such as shadows, reflections, refractions, and more, remains a relatively unexplored area. In terms of simulation, while the geometry of NeRF is implicit in its density field, it lacks a well-defined surface representation, making it difficult to detect and resolve collisions. Recent works have delved into enhancing the integration between NeRF and meshes, aiming to combine the photorealistic capabilities of NeRF with the versatility of meshes for rendering and simulation. Neural implicit surfaces [87, 80, 59, 19] are represented as learned Signed Distance Fields (SDF) within the NeRF framework. Meanwhile, methods like IRON [94] and NVDiffRec [52] extract explicit, textured meshes that are directly compatible with path tracing, offering practical benefits at the expense of a lossy discretization. Nerfstudio [72] renders NeRF and meshes separately, then composites the render passes with an occlusion mask. Unfortunately, this decoupled rendering approach offers no way to exploit the lighting and appearance information encoded in the NeRF volume to affect the rendered mesh appearance. Figure 6 visually compares our hybrid method to naively combining NeRF and surface rendering, and pure surface rendering.

We introduce a hybrid graphics pipeline that integrates the rendering and simulation of neural fields and meshes. For both representations, we consider lighting effects and contact handling for physical interaction. By unifying NeRF volume rendering and path tracing within the linear RGB space, we discover their Light Transport Equations exhibit similarities in terms of variables, forms, and principles. Leveraging their shared light transport behavior, we devise update rules for radiance and throughput variables, enabling seamless integration between NeRF and meshes. To incorporate shadows onto the NeRF, we employ differentiable surface rendering techniques [28] to estimate light sources and introduce secondary shadow rays during the ray marching process to determine visibility. Consequently, the NeRF rendering equation is modified to include a pointwise shadow mask.

For simulation, we adopt SDFs to represent geometry of neural fields, which is advantageous for physical contact handling and collision resolution. We then use position-based dynamics [42] for time integration. Our efficient hybrid rendering and simulation system is implemented in CUDA. To enhance usability, we have also incorporated

user-friendly Python interfaces. In summary, the key contributions of this work are:

- A two-way coupling between NeRF and surface representations for rendering and simulation.
- Integration with HDR data which can unify the color space of the path tracer and NeRF, with a strategy to estimate light sources and cast shadows on NeRF.
- An efficient rendering procedure that alternates ray marching and path tracing steps by blending the Light Transport Equations for both NeRF and meshes.
- An interactive, easy-to-use implementation with a high-level Python interface that connects the low-level rendering and simulation GPU kernels.

## 2. Related Work

### 2.1. Neural Fields and Surface Representations

Rendering of participating media has been extensively studied [57, 53] in classic graphics pipelines [58, 63, 50, 25]. In recent years, significant advancements have been made in this area [47, 55], yielding remarkable visual outcomes. Our work aims to expand upon this progress by incorporating the emergent Neural Radiance Fields (NeRF) [46], which have gained substantial popularity, into this exciting domain. Within the volume rendering framework, NeRF bakes the plenoptic function and volumetric density into spatial points. These points can be effectively parameterized by an MLP [46], convolutional networks [8], hash grid [49], point cloud [84], voxel [36, 70], or tensors [9]. It allows users to reconstruct photorealistic 3D static [71, 76, 66, 33, 81] or dynamic scenes [61, 83, 60, 20, 38] by casually capturing a few images [1]. Original NeRF takes seconds to render one single frame, while follow-up works have accelerated rendering speed [24, 78, 12, 21, 51, 88, 89, 3, 2, 79].

NeRF simplifies the creation of 3D content compared to classical mesh-based pipelines. However, addressing challenges of editing [93, 31] and decomposing the baked information [44, 68, 69, 85] is not trivial. Effort has been directed toward reconciling the advantages of both NeRF and surface-based paradigms. In rendering, [94] and [52] propose to use surface-based differentiable rendering to reconstruct textured meshes [23] from neural fields. Their reconstructed meshes can be imported to a surface rendering pipeline like Blender [15], but the original NeRF representation cannot be directly rendered with meshes. For simulation, NeRFEditing [92] proposes to use explicit mesh extracted by [80] to control the deformation of Neural Fields. Qiao et al. [65] further add full dynamics over the extracted tetrahedra mesh. Chu et al. [13] integrates the dynamics of smoke with neural fields. [14] also connects differentiable simulation to NeRF, where the density field and its gradient

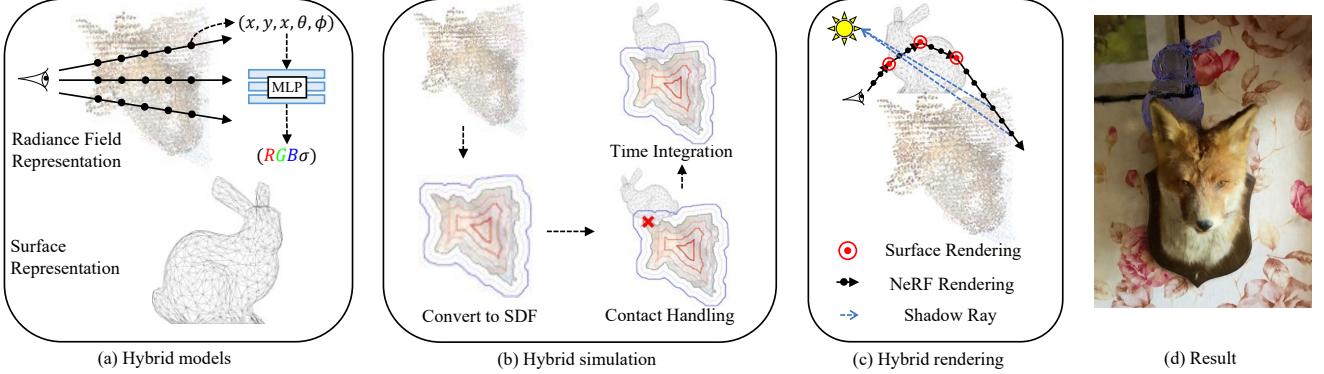


Figure 3: **Pipeline overview.** Our method takes an optimized radiance field model and surface meshes as inputs. We can run a physics simulation between the NeRF and meshes. The updated mesh vertices and NeRF transformations are synchronized to the renderer, which uses Monte Carlo simulation to sample ray paths. As the ray travels through space, it alternates between surface rendering (ray-tracing) and NeRF rendering (ray-marching), both updating its radiance.

are used to compute the contact. These methods aim to construct an end-to-end differentiable simulation and rendering pipeline, yet they have yet to couple the rendering.

## 2.2. Scene Editing and Synthesis

Our method enables inserting mesh assets into NeRF models of real-world captures. Editing of existing scenes is an active topic of study. For neural field representations, ray bending [75, 64, 34] is widely used to modify an optimized NeRF. It is possible to delete, add, duplicate, or actuate [10, 62, 82, 37] an area by bending the path of the rendering ray. [22] propose to train a NeRF for each object and compose them into a scene. ClimateNeRF [35] can change weather effects by modifying the density and radiance functions during ray marching. These methods study editing of isolated NeRF models. There are also inverse rendering works that decompose [86, 6] the information baked into NeRF, which can then be used to edit lighting and materials [32]. Such decomposition is useful, but assumes information like a priori knowledge of light sources, or synthetic scenes. They do not address inserting mesh into NeRF scenes. Besides NeRF, [30] inserts a virtual object into existing images by estimating the geometry and light source in the existing image. [11] insert vehicles in street scenes by warping textured cars using predicted 3D poses.

## 3. Method

In this section, we describe how radiance fields and polygonal meshes can be integrated into a unified rendering and simulation pipeline, an overview of which is visualized in Figure 3.

### 3.1. Rendering

NeRF can photorealistically reconstruct a 3D scene from a set of images, making it an appealing candidate for modeling environments and potentially valuable for traditional surface-based rendering. One possible approach to bridge these disparate representations is to render the meshes and NeRF volume in separate passes, and composite the results together in 2D image space. However, compositing in image space is susceptible to incorrect occlusion masks and inaccurate lighting. A more physically principled approach to this problem is identifying and exploiting the similarities in their respective light transport equations, which directly allows the radiance field and mesh to be incorporated in 3D space.

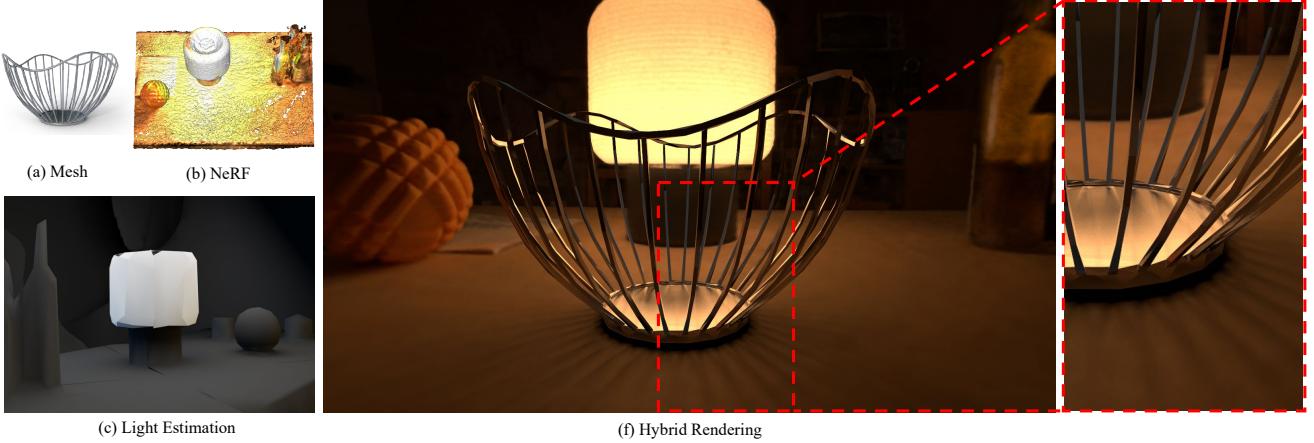
**Surface Rendering Equation.** The Light Transport Equation (LTE) for surface rendering is:

$$L(p, \omega_o) = L_e(p, \omega_o) + L_r(p, \omega_o) \quad (1)$$

$$L_r(p, \omega_o) = \int_{S^2} f_s(p, \omega_o, \omega_i) L_i(p, \omega_i) |\cos \theta_i| d\omega_i \quad (2)$$

$$L_i(p, \omega_i) = L(t(p, \omega_i), -\omega_i) \quad (3)$$

where  $p$  is a surface point;  $\omega_i, \omega_o$  are the directions of incident (incoming) and exitant (outgoing) radiance;  $S^2$  is the unit sphere sampling space for directions;  $L, L_e, L_i, L_r$  are the exitant, emitted, incident, and reflected radiance, respectively;  $\theta_i$  is the angle of incidence of illumination;  $f_s(p, \omega_o, \omega_i)$  is the bidirectional scattering distribution function (BSDF); and  $t(p, \omega)$  is the ray-casting function that computes the first surface intersected by the ray cast from  $p$  in the direction  $\omega$ .



**Figure 4: Shadow Casting.** We estimate the geometry and light source of the scene and insert a metal basket onto the desk. Our pipeline can render the realistic reflection and shadow effects caused by the synthetic mesh.

If a scene is represented solely by surfaces, the LTE in Equation 1 can be solved by Monte Carlo path tracing: for each pixel, a ray is randomly cast from the camera, its path constructed incrementally each time it hits, and bounces off of a surface. A natural but memory-inefficient way to implement this algorithm is to recursively compute Equation 1 and spawn a new ray upon each ray-surface intersection. Noticing that  $L_i(p, \omega_i)$  is independent of previous paths, the recursive process can be transformed into a weighted sum of radiance on each ray-surface intersection  $p_k$ . These weights  $T(p_k)$  are called *throughput*, and they depend on their predecessors' BSDF  $f_s(p, \omega_o, \omega_i)$ , illumination angle  $|\cos \theta_i|$ , and probability density function  $P$  of the scattering:

$$T(p_k) = \frac{T(p_{k-1}) \cdot f_s(p_k, \omega_{k-1}, \omega_k) |\cos(\theta_k)|}{P} \quad (4)$$

$T(p_k)$  and  $L_i(p_k, \omega)$  are the only variables essential to track and integrate for each bounce.

**Volumetric Rendering Equation.** The light transport equation for the volumetric medium is:

$$L(p, \omega_o) = \int_{t=0}^{t_f} \exp\left(-\int_{s=0}^t \sigma_t(p_s) ds\right) L_i(p_t, \omega_o) dt \quad (5)$$

$$L_i(p_t, \omega_o) = L_e(p_t, \omega_o) + L_s(p_t, \omega_o) \quad (6)$$

$$L_s(p_t, \omega_o) = \sigma_s(p_t) \int_{S^2} f_p(p_t, \omega_o, \omega_i) L(p_t, \omega_i) d\omega_i \quad (7)$$

where  $L_s$  is the (weighted) scattered radiance,  $\sigma_t$  and  $\sigma_s$  are the attenuations and scattering coefficients,  $f_p$  is the phase function of the volume, and  $p_s$  is on the ray  $p_s = p + s \cdot \omega_o$  (similar to  $p_t$ ). All other terms share the same definition as in surface rendering.

The integral in the volumetric LTE could again be solved using Monte Carlo methods. However, stochastic simulation of volumetric data is more challenging and expensive than surface data. A photon may change direction in a continuous medium, unlike the discrete bounces that occur only at surfaces. Therefore, rather than simulating the path of photons using Monte Carlo sampling, methods like NeRF [46] instead bake the attenuation coefficient  $\sigma(p) = \sigma_t(p)$  and view-dependent radiance  $r(p, \omega)$  onto each spatial point, and so there is no scattering. This circumvents solving Equation 7, thereby avoiding considering light transport, light sources, and material properties. Volume rendering under the NeRF formulation becomes:

$$r(p, \omega) = \int_0^{t_f} T(t) \sigma(p_t) r(p_t, \omega) dt, \quad (8)$$

$$T(t) = \exp\left(-\int_0^t \sigma(p_s) ds\right) \quad (9)$$

In Equation 8, the radiance and throughput are being updated, similar to surface rendering. However, note that the volumetric LTE denotes incident radiance at point  $p$  from direction  $\omega$  as  $L_i(p, \omega)$ , while NeRF denotes the same as  $r(p, \omega)$ . This terminology is indeed overloaded, as  $r(p, \omega)$  in the NeRF formula represents sRGB color, i.e. the result of applying a nonlinear tone-mapping function to the raw radiance value. The terms are related in that  $r(p, \omega) = \psi(L_i(p, \omega))$ , where  $\psi(\cdot)$  represents a tone-mapping function from linear to sRGB color space.

**Unifying color space of path tracing and ray marching.** The standard NeRF model accumulates sRGB color along rays cast from the camera into the volume: each point's color is represented by three 8-bit values, one for each color channel. Integrating these colors along the ray (weighted by transmittance) produces a final 8-bit color value, the ren-

dered pixel color, which is compared to the corresponding ground truth 8-bit pixel color to supervise NeRF training.

In contrast, path tracing assumes radiance values are expressed in linear color space. To relate NeRF and surface rendering in a physically meaningful way, they should ideally operate in a standard color space. To reconcile this difference, we train an HDR variant of NeRF, supervised with 32-bit HDR images directly rather than the standard 8-bit NeRF. The resulting HDR NeRF produces a 3-channel radiance in 32-bit *linear color space* at each sampled point. For details regarding HDR data acquisition, preprocessing, and HDR NeRF implementation, see Appendix E. As our focus here is to articulate the advantage of HDR NeRF in the context of the overall system, we forego a more general discussion of training NeRF in HDR and refer interested readers to [45] for a deep dive. The HDR NeRF rendering equation can thus be written as:  $L_i(p, \omega) = \int_0^{t_f} T(t)\sigma(p_t)L_i(p_t, \omega)dt$  where the transmittance term  $T(t)$  remains unchanged.

With this simple adjustment, the NeRF equation can now directly relate to the surface rendering equation. If capturing HDR training data is impractical, one can still use standard (LDR) NeRF. Since the NeRF volume acts as the only light source, and the total energy is dissipative during the light transport,  $L_i(p, \omega)$  will never exceed the NeRF volume's maximal radiance.  $\psi(\cdot)$  would then degenerate to an identity mapping such that  $L(p, \omega) = r(p, \omega)$ . In other words, in many practical cases, reasonable visual results could still be obtained if standard NeRF is used with our system, despite the resulting inaccuracy in the light transport simulation.

**Estimating light sources with differentiable surface rendering.** Sampling light sources for computing the shadow pass requires an approximate representation of the light sources (emitters) of the scene. Note that NeRF's volume rendering formulation bakes appearance into each point in the volume rather than simulating physically based light transport. To recover an explicit representation of light sources, we first reconstruct the scene's geometry as a neural SDF using *MonoSDF* [91], from which we extract an explicit mesh. Then, we employ a differentiable path tracer, *Mitsuba3* [28, 29], to estimate a UV Emission Texture for the mesh. We follow the general approach of [54], though we customize the optimization procedure since our goal is to estimate only the light sources, as opposed to a full BRDF estimation (more details about the optimization can be found in Appendix D).

Once the light source estimation has converged, we prune faces whose emission falls below a threshold from the explicit mesh, which is necessary for efficiency, as most faces in the explicit mesh do not emit light. The hybrid renderer then consumes the pruned mesh.

**Shadow rays.** We query additional rays during ray-

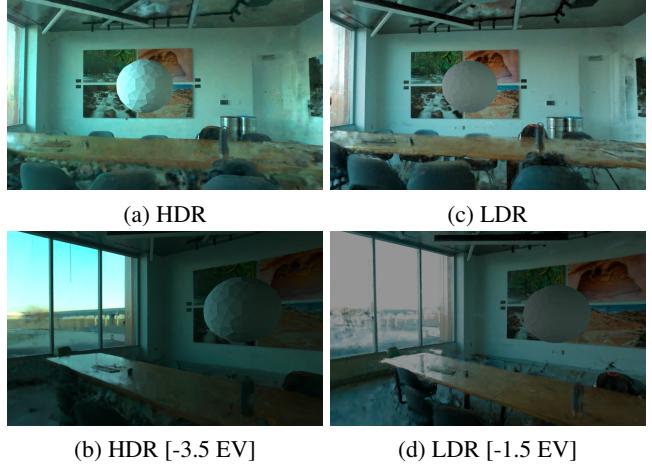


Figure 5: **HDR Volumetric Radiance Map.** The diffuse sphere is rendered using our hybrid algorithm. Images rendered using the NeRF trained in 32-bit HDR (a, b) achieve a higher level of lighting realism than those rendered with 8-bit LDR (c, d).

marching to cast shadows on NeRF. For each sampled point  $p_t$  in NeRF, we shoot a secondary ray from  $p_t$  to the light source (see the following subsection for details on estimating lighting sources). If an inserted mesh blocks this ray, then this  $p_t$  has a shadow mask  $m(p_t) = 1 - r_{src}$ , where  $r_{src}$  is the intensity of the light source. Non-blocked pixels have  $m_{shadow} = 1$ . The contribution of this point in Eq. 8 is then  $T(t)\sigma(p_t)m(p_t)r(p_t, \omega)$ . Fig. 4 shows that an inserted metal basket casts shadows on the NeRF desk.

**Hybrid Rendering Algorithm** Based on the Light Transport Equations mentioned above, we note that both surface and NeRF rendering integrate the throughput  $T(p)$  and radiance  $r(p, \omega) = L_i(p, \omega)$ . The differences are: (1) Surface rendering updates those values on discrete boundaries while NeRF accumulates them in the continuous space; (2)  $T(p)$  and  $r(p, \omega)$  are governed by the BSDF parameters in surface rendering, while by neural fields in NeRF. Therefore, we can alternate between the surface and NeRF rendering rules as they travel in space. Algorithm 1 is a summary of the hybrid rendering of NeRF and surface representations:

1. We use Monte Carlo path tracing to sample the ray-surface-intersections  $p_0 \rightarrow p_1 \rightarrow \dots \rightarrow p_n$ , where  $p_0$  is the camera center, and  $p_n$  is the termination of the path. At the beginning of the path, initialize accumulated throughput  $T(p_0) = 1$  and radiance  $r(p_0, \omega_0) = (0, 0, 0)$ . The termination conditions will be discussed in (4).
2. If shadows are needed, we estimate light source geometry and intensity  $r_{src}$  with differentiable surface rendering.

3. For each ray segment  $p_j \rightarrow p_{j+1}$ , we use the ray-marching algorithm to sample and integrate the NeRF medium. For the sampled points  $p_t$  on the ray, we shoot a ray from  $p_t$  to the light source (if any). If meshes block this point, set its *shadow mask* to be  $m(p_t) = 1 - r_{src}$  (or simply a constant close to 0), otherwise  $m(p_t) = 1$ . Then the throughput and *shadow masked* radiance between surface intersections  $p_j$  and  $p_{j+1}$  can be computed as,

$$T'(p_{j+1}) = T(p_j) \cdot \exp \left( - \int_{p_t} \sigma(p_t) dt \right) \quad (10)$$

$$L(p_{j+1}, \omega_j) = L(p_j, \omega_j) + \int_{p_t} T(t) \sigma(p_t) m(p_t) r(p_t, \omega) dt \quad (11)$$

where  $p_t \in (p_j, p_{j+1}]$  and

$$T(t) = T(p_j) \cdot \exp \left( - \int_0^t \sigma(p_s) ds \right)$$

is also accumulated from  $p_j$ .

4. At the end of a ray segment, we reach the interface  $p_{j+1}$  where the surface-rendering procedures occur. The direction  $\omega_{j+1}$  of the next ray is determined by sampling the BSDF, and the weighted illumination and emitted light at this point are added to the radiance:

$$L(p_{j+1}, \omega_{j+1}) = L(p_{j+1}, \omega_j) + T(p_j) L_e(p_{j+1}, \omega_j) \quad (12)$$

The throughput weight is updated as:

$$T(p_{j+1}) = \frac{T'(p_{j+1}) \cdot f_s(p_{j+1}, \omega_j, \omega_{j+1}) |\cos(\theta_{j+1})|}{P} \quad (13)$$

where  $P$  is the scattering probability density function.

5. In the end, the ray terminates at  $(p_e, \omega_e)$  if (1) it runs out of the scene; (2) current throughput  $T(p_e)$  is lower than a threshold; or (3) it meets the bounce limit.
6. As the rendering procedure is carried out over a linear 32-bit color space after the path tracing terminates for a given pixel, we can apply a nonlinear tone-mapping function, which we denote as  $\psi$ , to map from linear radiance to final sRGB color  $r(p_e, \omega_e)$  which is more suitable for displaying on a monitor:

$$r(p_e, \omega_e) = \psi(L(p_e, \omega_e)) \quad (14)$$

### 3.2. Simulation

We incorporate a dynamics simulator that supports rigid bodies, cloth, and deformable solids. Neural fields and meshes can be connected in the simulation pipeline by

---

#### Algorithm 1 Hybrid Rendering Pipeline

---

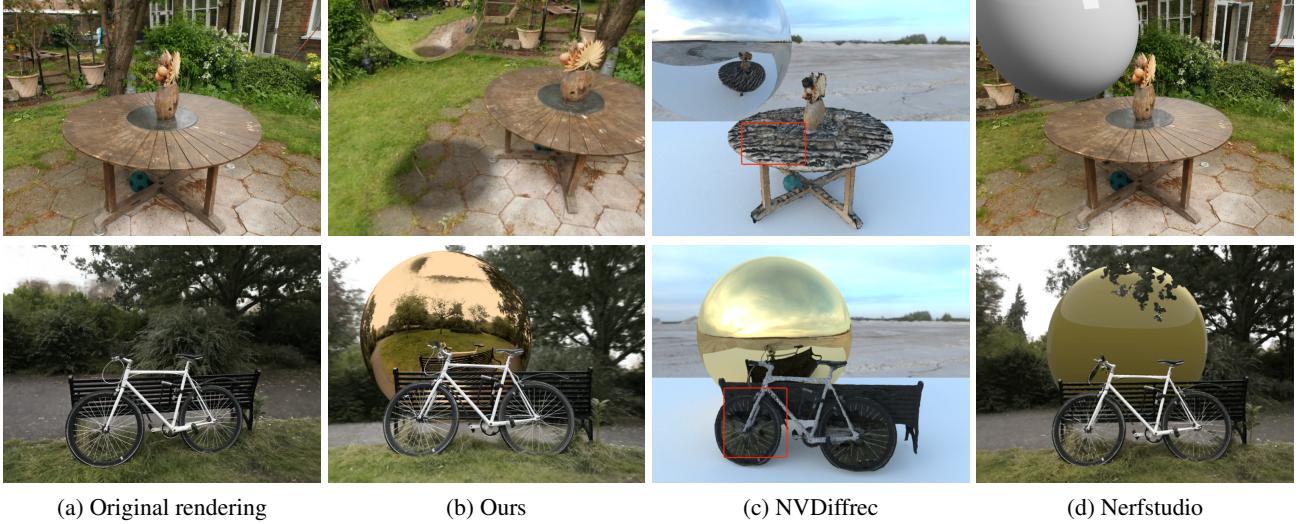
**Require:** Meshes and pretrained NeRF of the scene.  
Estimate light sources with differentiable surface rendering.  
**for** each pixel  $(u, v)$  in parallel **do**  
  Initialize  $p_0, \omega_0$  based on  $(u, v)$  and camera center.  
  Set throughput  $T(p_0) = 1$ .  
  Set radiance  $L(p_0, \omega_0) = (0, 0, 0)$ .  
  **for**  $j \in \{1, \dots, n_{bounces}\}$  **do**  
    Cast ray to find next intersection  $p_j$  (Eqn. 3).  
    March along ray  $p_{j-1} \rightarrow p_j$ .  
    Cast shadow rays to light sources.  
    Integrate  $T'(p_j)$  and  $L(p_j, \omega_{j-1})$  (Eqn. 10, 11).  
    Sample BSDF at  $p_j$  to get next ray direction  $\omega_j$   
    Update  $T(p_j)$  and  $L(p_j, \omega_j)$  (Eqn. 13, 12).  
    **break** if termination conditions satisfied.  
  **end for**  
  (*Path tracing endpoint denoted as  $(p_e, \omega_e)$* )  
  Apply tone-mapping function  $r = \psi(L(p_e, \omega_e))$ .  
**end for**

---

Signed Distance Fields (SDF) or reconstructed surface mesh. The SDF of the NeRF can be obtained in several ways during pre-processing. On the one hand, existing methods can directly learn the SDF, like NeuS [80] and VolSDF [87]. On the other hand, it is common to set a threshold value for the density field and extract the surface mesh through Marching Cube [40]. And the SDF can be converted from the mesh. The learned SDF has better quality but takes more time. We also implement an Instant-NGP version of NeuS (called NeuS-NGP) and accelerate the original code by more than 10 times. Users can make the trade-off depending on their needs.

We employ extended position-based dynamics (XPBD) [43, 48] to simulate the objects during runtime. We choose this dynamics model because it is fast and can support various physical properties. Collision detection is performed by querying the SDF of all vertices. All of these queries can be computed efficiently in parallel on a GPU. Given a detected collision from the SDF, we can also get the penetration depth and normal, which can be used to compute the contact forces.

In some scenarios, NeRF can represent movable objects (e.g. a scene can be a composition of several NeRF objects [73]) instead of a static background. We can get the homogenous transformation  $t \in \mathcal{R}^{4 \times 4}$  of the NeRF from simulation in each time step, which is used to inverse transform the homogenous coordinates  $t^{-1} \cdot p$  [64] when querying the color/density and sampling rays in the Instant-NGP hash grid. In Fig. 8 (b), we control a ball to interact with a NeRF chair [46] in real-time. The supplementary video further shows how the collision effect changes when the ball



**Figure 6: Rendering comparison for virtual object insertion.** We insert a reflective metal ball into the Garden and Bicycle scenes from the Mip-NeRF 360 dataset [5]. (b) Our hybrid method produces results of superior visual quality with fewer artifacts. (c) We extract the foreground mesh using NVDiffRec, then insert the synthetic ball and render using ray tracing. Extracting an explicit mesh object results in noticeable artifacts such as the noisy table surface, and missing thin structures like the bicycle’s wheel and the bench. (d) The 2D compositing workflow of Nerfstudio suffers from non-3D-aware occlusion masks and a limited ability to accommodate realistic interreflection.

and chair have different relative mass and velocities.

### 3.3. Implementation Details

The entire pipeline employs CUDA backends for computation and Python interfaces for interaction. For rendering, NeRF is trained with the default configuration using Instant-NGP [49]. We also implement an instant-NGP [49] version of NeuS [80] for efficient learning of the implicit SDF geometry. The path tracing algorithm is implemented using CUDA, embedded in Instant-NGP’s ray-marching procedures. We incorporate refractive, reflective, and Lambertian BSDF models. Physics simulation utilizes Warp [41], which just-in-time compiles Python code into CUDA kernels. The connection between rendering and simulation is facilitated by a Python interface using pybind11 [27]. Scene parameters can be easily created or modified through config files or Python APIs.

Our method can achieve a runtime of 1 to 40 frames per second, contingent upon the resolution, scene complexity, and dynamics. Figure 8 demonstrates a real-time game on a laptop that has been developed within our pipeline. The code for rendering, simulation, and fast SDF learning will be released as open-source software.

## 4. Experiments

### 4.1. Comparisons

**Rendering Comparisons.** In this section, we compare with other surface-based modeling and rendering methods in the

virtual object insertion task. Given a set of images, the traditional graphics pipeline would first reconstruct the surface. However, the 3D reconstruction step will usually introduce tremendous noise and errors. Our technique can directly render the virtual object in the photorealistic 3D scene without meshing the entire scene.

NVdiffrec [52] and IRON [94] are state-of-the-art textured mesh reconstruction methods. They combine neural fields and differentiable rendering methods to estimate the geometry and appearance of the objects from images. However, neither of them works on the full image because the topology of the background is too complex to optimize (e.g. the vegetation). We further provide the per-frame foreground mask to the comparison methods, and NVdiffRec can reconstruct foreground models.

In Figure 6 (c), we import the extracted mesh into Blender with a metallic ball and environment map, then render the scene with Cycles, a physically-based render. Both (b) ours and (c) NVDiffRec can model the reflection on the ball, but our object has better surface quality.

We also compare to Nerfstudio [72], a software library that allows users to train their own NeRF models. The user can export a camera path moving through the trained NeRF scene and a coarsely reconstructed mesh and import these into Blender [15] using the Nerfstudio Blender plugin. The user may then create their 3D content directly in Blender, using the imported camera path and mesh as a reference, and then render their 3D content, which is then composited in 2D over the rendered NeRF trajectory. However, while



(a) Soft body

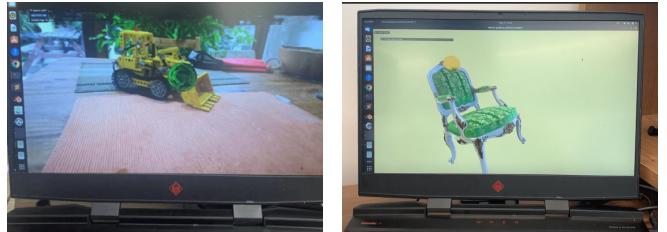
(b) Cloth

**Figure 7: Qualitative results for simulation.** In addition to rigid body simulation (Fig. 8), our method can also simulate soft bodies. (a) is a twisting Neo-Hookean FEM mesh. (b) is a thin shell cloth covering the fox. Please see the supplementary video for more simulation results.

the coarse mesh may be used to generate a visibility mask to produce occlusion, this approach is limited in that the coarsely reconstructed mesh is likely not accurate enough to provide clean-looking occlusion results, more complex occlusion situations such as occluding objects at multiple layers of depth will be challenging to generate individual visibility masks, and this requires manual effort to composite the results. Appendix F includes some more quantitative comparisons with those methods. Moreover, LumaAI recently released a closed-source UE plugin, and we run a comparison against it.

**Rendering with HDR NeRF.** We provide a qualitative comparison between rendering using the standard (LDR) NeRF and its HDR counterpart, shown in Figure 5. Notice that in the images rendered with the HDR model, the lighting cast from the environment onto the mesh appears much more faithful to the scene’s true intensity (and therefore directionality). This is also not surprising, considering the longstanding use of High Dynamic Range Image-Based Lighting in the traditional graphics pipeline [16]. HDR images are often created by recovering the unknown nonlinear tone-mapping function from a series of bracketed-exposure duration, then using the inverse function to map the images back to linear color space, and merging them into a single 32-bit result, as introduced by Debevec and Malik [17]. As a result, HDR images are much better suited to capture a scene’s full range of absolute and relative radiance values and avoid highly lossy clipping, which can be especially problematic in very bright parts of an image. HDR NeRF may therefore be interpreted as a volumetric HDR lighting map.

Importantly, our hybrid rendering algorithm allows us to utilize such an HDR volumetric radiance map fully. As an HDR radiance map is particularly useful for representing the indirect bounce lighting of the scene, future work on extending such learned lighting models to direct lighting (e.g. directional or point sources that cast hard shadows) would



(a) Static NeRF background (b) NeRF (chair) as dynamic object

**Figure 8: Real-time photorealistic gaming on a laptop.** Our rendering and simulation engine can be interactive and run in real-time on a laptop with an NVIDIA GeForce RTX 2070 Max-Q GPU. In this example, a user can control the motion of the synthetic glass ball and interact with the background (collision and light effect).

be a promising direction. Our hybrid rendering algorithm is crucial to enable such an investigation.

**Simulation Comparisons.** Besides static scenes, our pipeline can also simulate dynamic scenes with SDF-based contact handling. Appendix B shows that the SDF-based representation has better collision handling than vanilla NeRF density fields. We also compare with NeRF-based simulation [14, 65] in Appendix C and ours achieves better performance.

## 4.2. Performance

In this section, we will show how our method can be used in photorealistic real-time gaming and physically-based simulation. A detailed profiling of our rendering and simulation modules can also be found in Appendix A.

**Photorealistic Real-time Gaming.** In addition to photorealistic rendering, our pipeline is fast, aiming to serve as a real-time neural-fields game engine. As shown in Figure 8, we have implemented an interactive game where players can control the ball’s motion using a keyboard and adjust camera viewing angles using a mouse. The background scene, excluding the green ball, is modeled by NeRF. The ball can have contact with the table and bulldozer. Players can also observe the bending of rays as they pass through the refractive glass ball. Supplementary material includes a recording of the real-time game. The game runs on a laptop with an NVIDIA GeForce RTX 2070 Max-Q GPU. Through our pipeline, game developers can seamlessly integrate animatable objects with photorealistic NeRFs.

**Qualitative Results.** Our method can simulate different types of dynamics with the support of Warp [41]. In Figure 7, the background fox is modeled as NeRF. (a) is a cube modeled by Neo-Hookean FEM mesh, where we can also see the changing light effect as it is twisted. (b) is a piece of cloth falling down to the fox. Our pipeline can handle the

collision and light effects of such thin shells.

### 4.3. Applications

Our methods could be applied to many situations, where NeRFs can improve the realism of synthetic scenes.

**Driving simulation** is important for developing, training, and testing autonomous driving systems. With the large number of images captured around roads, people can train NeRF for street views [71]. With our method, people can set up the driving simulation inside those Photorealistic NeRFs (see Figure 9 (a)) and insert synthetic vehicles [4, 74]. Such realistic virtual environments can help minimize the sim-to-real gap in self-driving cars.

**Room layout design** can help users design their homes and purchase furniture. After taking pictures and building a NeRF model for their room, customers can shop furniture [26, 67] virtually and design the room layout as shown in Figure 9 (b).

**Virtual try-on** using our methods can dynamically simulate the cloth on a human body captured by NeRF. In Figure 9 (c), we place a cloak on the human body and simulate how it swings in the wind.

**Digital human** applications are one of the key interests in VR/AR and the metaverse. With our methods, users can easily collect and build their virtual world by NeRF and then render their human-body model in that scene. This could be useful for movie making, webcasting, virtual performance, cyber-tourism, etc. Figure 9 (d) renders a futuristic ‘mercury man’ [39] jumping in the park using NeRF.

## 5. Conclusion

In summary, motivated by integrating NeRF into the traditional graphics pipeline, our hybrid rendering method can render dynamically changing meshes in a photo-realistic NeRF environment, without costly surface reconstruction. We also equip the resulting renderer with a simulator, making it suitable as a real-time NeRF-based game engine.

There are some limitations in this work. (1) The currently implemented renderer cannot cast shadows and illumination on NeRF points. Decomposing NeRFs can make the relighting more realistic. (2) Our renderer offers basic, essential functions; support for environment maps, UV maps, and image textures for higher rendering quality can be a natural extension. (3) Additional interfaces can also enable users to take advantage of more mature infrastructures if integrated into more widely-used platforms, e.g. Blender, Unreal, etc. (4) We can further improve the runtime performance and integrate the pipeline with larger-scale NeRFs.

**Acknowledgements.** This research is supported in part by Dr. Barry Mersky and Capital One E-Nnovate Endowed Professorships, and ARL Cooperative Agreement

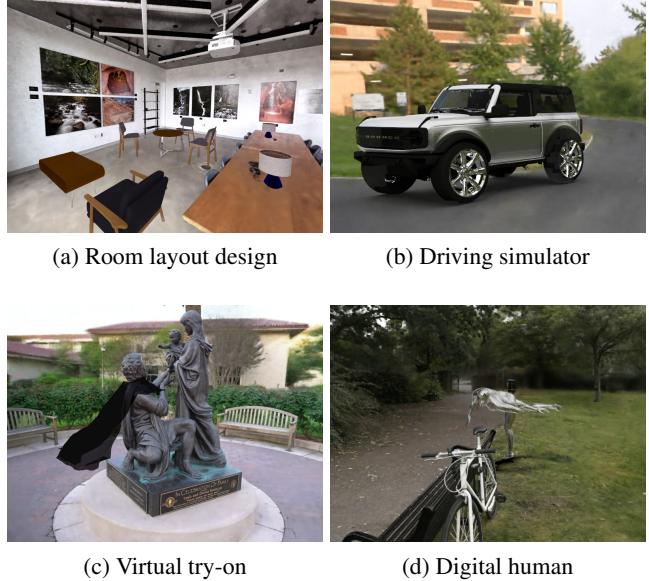


Figure 9: **Application of mixing NeRF with meshes.** Our method can be used for realistic driving simulation, room layout design, virtual try-on, and digital humans.

W911NF2120076. Yi-Ling would also like to thank the support from Meta Fellowship and Dr. Chunsheng Hu’s assistance in drawing the diagram.

## References

- [1] Luma AI. Luma ai. <https://lumalabs.ai/>, 2022. 2
- [2] Benjamin Attal, Jia-Bin Huang, Christian Richardt, Michael Zollhoefer, Johannes Kopf, Matthew O’Toole, and Changil Kim. Hyperreal: High-fidelity 6-dof video with ray-conditioned sampling. In *CVPR*, 2023. 2
- [3] Benjamin Attal, Jia-Bin Huang, Michael Zollhöfer, Johannes Kopf, and Changil Kim. Learning neural light fields with ray-space embedding. In *CVPR*, 2022. 2
- [4] baehappy1322. The all new ford bronco. 9
- [5] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. *CVPR*, 2022. 1, 7, 13
- [6] Mark Boss, Raphael Braun, Varun Jampani, Jonathan T. Barron, Ce Liu, and Hendrik P.A. Lensch. Nerd: Neural reflectance decomposition from image collections. In *ICCV*, 2021. 3
- [7] Brent Burley. Physically-based shading at disney. 2012. 13
- [8] Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3D generative adversarial networks. In *arXiv*, 2021. 2
- [9] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorrf: Tensorial radiance fields. In *European Conference on Computer Vision (ECCV)*, 2022. 2

- [10] Jianchuan Chen, Ying Zhang, Di Kang, Xuefei Zhe, Linchao Bao, Xu Jia, and Huchuan Lu. Animatable neural radiance fields from monocular rgb videos, 2021. 3
- [11] Yun Chen, Frieda Rong, Shivam Duggal, Shenlong Wang, Xinchen Yan, Sivabalan Manivasagam, Shangjie Xue, Ersin Yumer, and Raquel Urtasun. Geosim: Realistic video simulation via geometry-aware composition for self-driving. In *CVPR*, 2021. 3
- [12] Zhiqin Chen, Thomas Funkhouser, Peter Hedman, and Andrea Tagliasacchi. Mobilenerf: Exploiting the polygon rasterization pipeline for efficient neural field rendering on mobile architectures. *arXiv preprint arXiv:2208.00277*, 2022. 2
- [13] Mengyu Chu, Lingjie Liu, Quan Zheng, Erik Franz, Hans-Peter Seidel, Christian Theobalt, and Rhaleb Zayer. Physics informed neural fields for smoke reconstruction with sparse data. *ACM Transactions on Graphics*, 2022. 2
- [14] Simon Cleac'h, Hong-Xing Yu, Michelle Guo, Taylor Howell, Ruohan Gao, Jiajun Wu, Zac Manchester, and Mac Schwager. Differentiable physics simulation of dynamics-augmented neural objects, 10 2022. 2, 8, 13
- [15] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. 2, 7
- [16] Paul Debevec. Rendering synthetic objects into real scenes: Bridging traditional and image-based graphics with global illumination and high dynamic range photography. *SIGGRAPH*, 1998. 8
- [17] Paul E. Debevec and Jitendra Malik. Recovering high dynamic range radiance maps from photographs. *SIGGRAPH* 97, 1997. 8, 15
- [18] Tao Du, Kui Wu, Pingchuan Ma, Sebastien Wah, Andrew Spielberg, Daniela Rus, and Wojciech Matusik. DiffPD: Differentiable projective dynamics with contact. *arXiv:2101.05917*, 2021. 13
- [19] Stefano Esposito, Daniele Baieri, Stefan Zellmann, André Hinkenjann, and Emanuele Rodolà. Kiloneus: Implicit neural representations with real-time global illumination. *arXiv preprint arXiv:2206.10885*, 2022. 2
- [20] Chen Gao, Ayush Saraf, Johannes Kopf, and Jia-Bin Huang. Dynamic view synthesis from dynamic monocular video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 2
- [21] Stephan J Garbin, Marek Kowalski, Matthew Johnson, Jamie Shotton, and Julien Valentin. Fastnerf: High-fidelity neural rendering at 200fps. In *ICCV*, 2021. 2
- [22] Michelle Guo, Alireza Fathi, Jiajun Wu, and Thomas Funkhouser. Object-centric neural scene rendering, 2022. 3
- [23] Jon Hasselgren, Nikolai Hofmann, and Jacob Munkberg. Shape, Light, and Material Decomposition from Images using Monte Carlo Rendering and Denoising. *arXiv:2206.03380*, 2022. 2
- [24] Peter Hedman, Pratul P Srinivasan, Ben Mildenhall, Jonathan T Barron, and Paul Debevec. Baking neural radiance fields for real-time view synthesis. In *ICCV*, 2021. 2
- [25] Binh-Son Hua, Adrien Gruson, Victor Petitjean, Matthias Zwicker, Derek Nowrouzezahrai, Elmar Eisemann, and Toshiya Hachisuka. A survey on gradient-domain rendering. In *Computer Graphics Forum*, volume 38, pages 455–472. Wiley Online Library, 2019. 2
- [26] Ivo. High-poly modern wood chair. 9
- [27] Wenzel Jakob, Jason Rhinelander, and Dean Moldovan. pybind11 — seamless operability between c++11 and python, 2016. <https://github.com/pybind/pybind11>. 7
- [28] Wenzel Jakob, Sébastien Speierer, Nicolas Roussel, Merlin Nimier-David, Delio Vicini, Tizian Zeltner, Baptiste Nicolet, Miguel Crespo, Vincent Leroy, and Ziyi Zhang. Mitsuba 3 renderer, 2022. <https://mitsuba-renderer.org>. 2, 5, 13
- [29] Wenzel Jakob, Sébastien Speierer, Nicolas Roussel, and Delio Vicini. Drjit: A just-in-time compiler for differentiable rendering. *Transactions on Graphics (Proceedings of SIGGRAPH)*, 41(4), July 2022. 5, 13
- [30] Kevin Karsch, Varsha Hedau, David Forsyth, and Derek Hoiem. Rendering synthetic objects into legacy photographs. *ACM Trans. Graph.*, 30(6):1–12, dec 2011. 3
- [31] Yoni Kasten, Dolev Ofri, Oliver Wang, and Tali Dekel. Layered neural atlases for consistent video editing. *ACM Transactions on Graphics (TOG)*, 40(6):1–12, 2021. 2
- [32] Sosuke Kobayashi, Eiichi Matsumoto, and Vincent Sitzmann. Decomposing nerf for editing via feature field distillation. In *Advances in Neural Information Processing Systems*, volume 35, 2022. 3
- [33] Georgios Kopanas, Thomas Leimkühler, Gilles Rainer, Clément Jambon, and George Drettakis. Neural point catacaustics for novel-view synthesis of reflections. *ACM Transactions on Graphics*, 41(6):Article–201, 2022. 2
- [34] Xuan Li, Yi-Ling Qiao, Peter Yichen Chen, Krishna Murthy Jatavallabhula, Ming Lin, Chenfanfu Jiang, and Chuang Gan. Pac-nerf: Physics augmented continuum neural radiance fields for geometry-agnostic system identification. *ICLR*, 2023. 3
- [35] Yuan Li, Zhi-Hao Lin, David Forsyth, Jia-Bin Huang, and Shenlong Wang. Climatenerf: Physically-based neural rendering for extreme climate synthesis. *arXiv e-prints*, pages arXiv–2211, 2022. 3
- [36] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. Neural sparse voxel fields. *NeurIPS*, 2020. 2
- [37] Lingjie Liu, Marc Habermann, Viktor Rudnev, Kripasindhu Sarkar, Jiatao Gu, and Christian Theobalt. Neural actor: Neural free-view synthesis of human actors with pose control. *ACM Trans. Graph.(ACM SIGGRAPH Asia)*, 2021. 3
- [38] Yu-Lun Liu, Chen Gao, Andreas Meuleman, Hung-Yu Tseng, Ayush Saraf, Changil Kim, Yung-Yu Chuang, Johannes Kopf, and Jia-Bin Huang. Robust dynamic radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2
- [39] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, Oct. 2015. 9

- [40] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. *ACM siggraph computer graphics*, 21(4):163–169, 1987. 6
- [41] Miles Macklin. Warp: A high-performance python framework for gpu simulation and graphics. <https://github.com/nvidia/warp>, March 2022. NVIDIA GPU Technology Conference (GTC). 7, 8
- [42] Miles Macklin, Matthias Müller, and Nuttapong Chentanez. Xpbd: Position-based simulation of compliant constrained dynamics. MIG ’16, page 49–54. Association for Computing Machinery, 2016. 2
- [43] Miles Macklin, Matthias Müller, and Nuttapong Chentanez. Xpbd: position-based simulation of compliant constrained dynamics. In *Proceedings of the 9th International Conference on Motion in Games*, pages 49–54, 2016. 6
- [44] Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *CVPR*, 2021. 2
- [45] 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. 5
- [46] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 1, 2, 4, 6
- [47] Bailey Miller, Iliyan Georgiev, and Wojciech Jarosz. A null-scattering path integral formulation of light transport. *ACM Transactions on Graphics (Proceedings of SIGGRAPH)*, 38(4), July 2019. 2
- [48] Matthias Müller, Miles Macklin, Nuttapong Chentanez, Stefan Jeschke, and Tae-Yong Kim. Detailed rigid body simulation with extended position based dynamics. In *Computer Graphics Forum*, volume 39, pages 101–112. Wiley Online Library, 2020. 6
- [49] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4):102:1–102:15, July 2022. 1, 2, 7, 13
- [50] Thomas Müller, Marios Papas, Markus Gross, Wojciech Jarosz, and Jan Novák. Efficient rendering of heterogeneous polydisperse granular media. *ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH Asia 2016)*, 35(6):168:1–168:14, 2016. 2
- [51] Thomas Müller, Fabrice Rousselle, Jan Novák, and Alexander Keller. Real-time neural radiance caching for path tracing. *ACM Trans. Graph.*, 40(4):36:1–36:16, Aug. 2021. 2
- [52] Jacob Munkberg, Jon Hasselgren, Tianchang Shen, Jun Gao, Wenzheng Chen, Alex Evans, Thomas Mueller, and Sanja Fidler. Extracting Triangular 3D Models, Materials, and Lighting From Images. *arXiv:2111.12503*, 2021. 2, 7
- [53] Thomas Müller, Marios Papas, Markus Gross, Wojciech Jarosz, and Jan Novák. Efficient rendering of heterogeneous polydisperse granular media. *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia)*, 35(6):168:1–168:14, Dec. 2016. 1, 2
- [54] Merlin Nimier-David, Zhao Dong, Wenzel Jakob, and Anton Kaplanyan. Material and Lighting Reconstruction for Complex Indoor Scenes with Texture-space Differentiable Rendering. In Adrien Bousseau and Morgan McGuire, editors, *Eurographics Symposium on Rendering - DL-only Track*. The Eurographics Association, 2021. 5, 13
- [55] Merlin Nimier-David, Thomas Müller, Alexander Keller, and Wenzel Jakob. Unbiased inverse volume rendering with differential trackers. *ACM Trans. Graph.* 2
- [56] Jan Novák. *Efficient Many-Light Rendering of Scenes with Participating Media*. PhD thesis, Karlsruhe Institute of Technology”, May 2014. 1
- [57] Jan Novák, Iliyan Georgiev, Johannes Hanika, and Wojciech Jarosz. Monte Carlo methods for volumetric light transport simulation. *Computer Graphics Forum (Proceedings of Eurographics - State of the Art Reports)*, 37(2), May 2018. 2
- [58] Jan Novák, Iliyan Georgiev, Johannes Hanika, Jaroslav Krivánek, and Wojciech Jarosz. Monte carlo methods for physically based volume rendering. In *ACM SIGGRAPH 2018 Courses*, SIGGRAPH ’18, pages 14:1–14:1, New York, NY, USA, 2018. ACM. 2
- [59] Michael Oechsle, Songyou Peng, and Andreas Geiger. Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *International Conference on Computer Vision (ICCV)*, 2021. 2
- [60] Keunhong Park, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Steven M. Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. *ICCV*, 2021. 2
- [61] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M. Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *ACM Trans. Graph.*, 40(6), dec 2021. 2
- [62] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *CVPR*, 2021. 3
- [63] Matt Pharr, Wenzel Jakob, and Greg Humphreys. *Physically based rendering: From theory to implementation*. Morgan Kaufmann, 2016. 2
- [64] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. *arXiv preprint arXiv:2011.13961*, 2020. 3, 6
- [65] Yi-Ling Qiao, Alexander Gao, and Ming C. Lin. Neu-physics: Editable neural geometry and physics from monocular videos. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2022. 2, 8, 13, 16
- [66] Konstantinos Rematas, Andrew Liu, Pratul P Srinivasan, Jonathan T Barron, Andrea Tagliasacchi, Thomas Funkhouser, and Vittorio Ferrari. Urban radiance fields. In *CVPR*, 2022. 2
- [67] robula72. coffee table marina. 9
- [68] Viktor Rudnev, Mohamed Elgharib, William Smith, Lingjie Liu, Vladislav Golyanik, and Christian Theobalt. Neural ra

- diance fields for outdoor scene relighting. *arXiv preprint arXiv:2112.05140*, 2021. 2
- [69] Pratul P Srinivasan, Boyang Deng, Xiuming Zhang, Matthew Tancik, Ben Mildenhall, and Jonathan T Barron. Nerv: Neural reflectance and visibility fields for relighting and view synthesis. In *CVPR*, 2021. 2
- [70] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. In *CVPR*, 2022. 2
- [71] Matthew Tancik, Vincent Casser, Xinchen Yan, Sabeek Pradhan, Ben Mildenhall, Pratul P Srinivasan, Jonathan T Barron, and Henrik Kretzschmar. Block-nerf: Scalable large scene neural view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8248–8258, 2022. 2, 9
- [72] Matthew Tancik\*, Ethan Weber\*, Evonne Ng\*, Ruilong Li, Brent Yi, Terrance Wang, Alexander Kristoffersen, Jake Austin, Kamyar Salahi, Abhik Ahuja, David McAllister, and Angjoo Kanazawa. Nerfstudio: A framework for neural radiance field development, 2022. 2, 7, 13
- [73] Jiaxiang Tang, Xiaokang Chen, Jingbo Wang, and Gang Zeng. Compressible-composable nerf via rank-residual decomposition. *arXiv preprint arXiv:2205.14870*, 2022. 6
- [74] toivo. Car scene. 9
- [75] Edgar Treitschke, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Christoph Lassner, and Christian Theobalt. Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. In *IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2021. 3
- [76] Haithem Turki, Deva Ramanan, and Mahadev Satyanarayanan. Mega-nerf: Scalable construction of large-scale nerfs for virtual fly-throughs. In *CVPR*, 2022. 2
- [77] Delio Vicini, Sébastien Speirer, and Wenzel Jakob. Path replay backpropagation: Differentiating light paths using constant memory and linear time. *Transactions on Graphics (Proceedings of SIGGRAPH)*, 40(4):108:1–108:14, Aug. 2021. 13
- [78] Huan Wang, Jian Ren, Zeng Huang, Kyle Olszewski, Menglei Chai, Yun Fu, and Sergey Tulyakov. R2l: Distilling neural radiance field to neural light field for efficient novel view synthesis. *arXiv preprint arXiv:2203.17261*, 2022. 2
- [79] Liao Wang, Jiakai Zhang, Xinhang Liu, Fuqiang Zhao, Yan-shun Zhang, Yingliang Zhang, Minye Wu, Jingyi Yu, and Lan Xu. Fourier plenoctrees for dynamic radiance field rendering in real-time. In *CVPR*, 2022. 2
- [80] 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. *NeurIPS*, 2021. 2, 6, 7, 13
- [81] Ziyu Wang, Liao Wang, Fuqiang Zhao, Minye Wu, Lan Xu, and Jingyi Yu. Mirrornarf: One-shot neural portrait radiance field from multi-mirror catadioptric imaging. In *2021 IEEE International Conference on Computational Photography (ICCP)*, pages 1–12. IEEE, 2021. 2
- [82] Chung-Yi Weng, Brian Curless, Pratul P. Srinivasan, Jonathan T. Barron, and Ira Kemelmacher-Shlizerman. HumanNeRF: Free-viewpoint rendering of moving people from monocular video. In *CVPR*, pages 16210–16220, June 2022. 3
- [83] Wenqi Xian, Jia-Bin Huang, Johannes Kopf, and Changil Kim. Space-time neural irradiance fields for free-viewpoint video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9421–9431, 2021. 2
- [84] Qiangeng Xu, Zexiang Xu, Julien Philip, Sai Bi, Zhixin Shu, Kalyan Sunkavalli, and Ulrich Neumann. Pointnerf: Point-based neural radiance fields. *arXiv preprint arXiv:2201.08845*, 2022. 2
- [85] Bangbang Yang, Yinda Zhang, Yinghao Xu, Yijin Li, Han Zhou, Hujun Bao, Guofeng Zhang, and Zhaopeng Cui. Learning object-compositional neural radiance field for editable scene rendering. In *ICCV*, 2021. 2
- [86] Yao Yao, Jingyang Zhang, Jingbo Liu, Yihang Qu, Tian Fang, David McKinnon, Yanghai Tsin, and Long Quan. Neilf: Neural incident light field for physically-based material estimation. In *ECCV*, page 700–716, Berlin, Heidelberg, 2022. Springer-Verlag. 3
- [87] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. In *Neurips*, 2021. 2, 6
- [88] Alex Yu, Sara Fridovich-Keil, Matthew Tancik, Qinrong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxtels: Radiance fields without neural networks. *arXiv preprint arXiv:2112.05131*, 2021. 2
- [89] Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. Plenotrees for real-time rendering of neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5752–5761, 2021. 2
- [90] Zehao Yu, Anpei Chen, Bozidar Antic, Songyou Peng Peng, Apratim Bhattacharyya, Michael Niemeyer, Siyu Tang, Torsten Sattler, and Andreas Geiger. Sdfstudio: A unified framework for surface reconstruction, 2022. 13
- [91] 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. 5, 13
- [92] Yu-Jie Yuan, Yang-Tian Sun, Yu-Kun Lai, Yuewen Ma, Rongfei Jia, and Lin Gao. Nerf-editing: Geometry editing of neural radiance fields. In *Computer Vision and Pattern Recognition (CVPR)*, 2022. 2
- [93] Jiakai Zhang, Xinhang Liu, Xinyi Ye, Fuqiang Zhao, Yan-shun Zhang, Minye Wu, Yingliang Zhang, Lan Xu, and Jingyi Yu. Editable free-viewpoint video using a layered neural representation. *ACM Transactions on Graphics (TOG)*, 40(4):1–18, 2021. 2
- [94] Kai Zhang, Fujun Luan, Zhengqi Li, and Noah Snavely. Iron: Inverse rendering by optimizing neural sdf’s and materials from photometric images. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022. 2, 7

## A. Runtime and Memory

To break down the time and memory cost of our method, Figure 12 profiles the simulation and rendering modules as the number of synthetic objects increases. In the test scene, we throw  $n \in \{1, 10, 100, 1000, 10000\}$  balls onto a Fox [49] represented by NeRF. The rendering resolution is  $900 \times 450$ .

As seen in the figure, the simulation and rendering memory stay constant at a low level (0.5 GB and 1.0 GB, respectively) throughout the entire experiment. The run time for both modules scales practically linearly w.r.t. the number of objects.

Moreover, the simulation module is independent of the resolution, and the rendering time scales linearly w.r.t. the number of pixels (see left). This pipeline can run in real time (20 FPS) at  $600 \times 300$  resolution.

## B. Comparison with density fields

We compare the simulation quality with different collision geometry. In Figure 14, we drop four balls above a messy counter top [5]. In this simulation scene, (a) models the collision geometry using learned SDF by our NeuS-NGP and (b) directly uses Marching-Cube mesh from instant-NGP density fields (with resolution  $256 \times 256 \times 256$ ). Two of the balls in (b) ‘sink’ into the countertop, which is unrealistic. In contrast, our SDF-based simulation in (a) can correctly simulate the interaction between the balls and the background objects. More dynamics results can be found in our supplementary video.

## C. Comparison with other NeRF-based simulation.

There are two recent works [14, 65] that can simulate in neural fields. [14] establish a collision model based on the density field. In our simulation, we found that the density fields are usually noisy and inadequate to model a surface well for accurate contact processing. NeuPhysics[65] aims at differentiable simulation and rendering, and it extracts hexahedra mesh from learned SDF [80] and simulates the mesh using [18]. NeuPhysics can only simulate existing NeRF objects instead of synthetic objects. Table 13 shows that our method runs at least an order of magnitude faster when simulating a bouncing ball.

## D. Light Source Estimation - Implementation Details

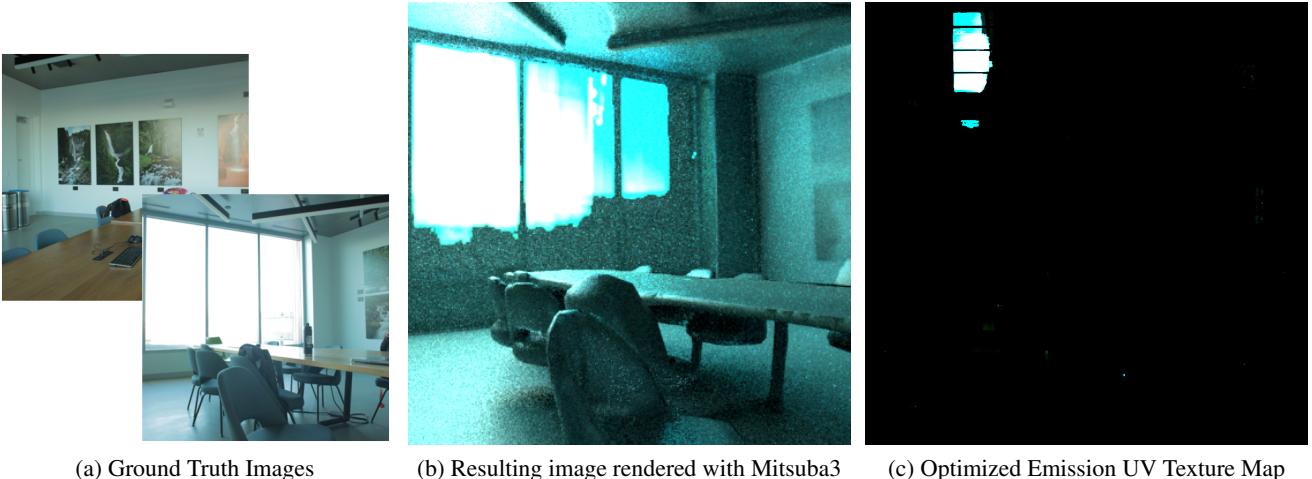
### D.1. Geometry Reconstruction

Given our choice to represent light sources as area lights defined on an explicit mesh, we first need to reconstruct the geometry of the scene. We choose to use MonoSDF [91], which leverages monocular geometric priors (depth and normal estimation) in the neural implicit surface reconstruction process. Importantly, these priors especially help in reconstructing surfaces that have only a low degree of visual texture, which is very common for walls and flat tabletop surfaces. For differentiable surface rendering, having reasonably accurate geometry is critical for achieving good convergence, as noisy geometry will tend to bias the result to bad local minima. Empirically, MonoSDF achieves sufficiently accurate geometry for this purpose. Once we have an optimized signed distance field, we convert it to an explicit mesh using Marching Cubes, then UV unwrap the mesh to obtain a UV texture map, using SDFStudio [90, 72]. This step also results in a UV texture map whose values are determined by querying the underlying MonoSDF appearance model, which we use as an albedo UV texture map. The UV-unwrapped mesh is the input to the differentiable surface render, described below.

### D.2. Differentiable Surface Rendering

We implement our differentiable surface rendering optimization using Mitsuba3 [28], which is built on Dr. Jit [29], a just-in-time compiler that is specialized for rendering use cases. Our inverse rendering procedure is generally based on [54], though our specific implementation details are as follows. The UV-unwrapped scene mesh is assigned a Principled BSDF material [7], as well as an emission UV texture map, which constitute the inputs to the differentiable renderer. The emission UV texture map,  $T_{\text{emission}}$ , with dimensions  $(h_{tx}, w_{tx}, c_{tx})$ , is the variable that we are interested in optimizing. It is initialized uniformly with near-zero values. The other BSDF parameters remain fixed (albedo texture map, specular transmission map with default value 1.0, and roughness map with default value 0.5). For rendering, we use a Path-Replay Backpropagation (PRB) integrator [77], and limit the light transport simulation to a maximum depth of 3 (2 bounces). This stands in contrast to differentiable rasterization approaches to inverse rendering, which are limited to a maximum depth of 2 (1 bounce), and are therefore unable to account for indirect lighting in the optimization. We also directly render emitters (as opposed to only emitting light onto surfaces without being directly visible).

Given a set of ground-truth 32-bit HDR images  $\{I_1, \dots, I_n\}$  with dimensions  $(h_{im}, w_{im}, c_{im})$  and known camera poses, our primary objective is to minimize the rendering loss (mean-squared error) between the ground truth images, and images



(a) Ground Truth Images      (b) Resulting image rendered with Mitsuba3      (c) Optimized Emission UV Texture Map

Figure 10: **Optimizing Emission UV Texture Map.** We optimize the (c) Emission UV Texture Map by minimizing a rendering loss objective between (a) ground truth images and (b) images rendered with Mitsuba3, a differentiable surface rendering engine.

rendered from identical camera poses. Since the variables that we are optimizing are the emission UV texture map values, we also find that applying uniform L1 regularization to the optimization variable corresponds well to the inductive bias that the majority of the scene does not emit light, and therefore the emission UV texture map should be encouraged to be sparse. For each optimization step  $i \in [1, n]$  in a given optimization epoch, our objective is therefore to minimize the following, where  $\alpha$  is a tunable hyperparameter:

$$\text{minimize } \frac{1}{(h_{im} * w_{im} * c_{im})} \|I_i - \hat{I}_i\|^2 + \frac{\alpha}{(h_{tx} * w_{tx} * c_{tx})} |T_{emission}|$$

In addition to the rendering loss and L1 regularization in texture space, we also find that periodically clipping low values and boosting large values in the emission texture space encourages convergence toward correct results. The motivation is that periodically during the optimization procedure (according to a tunable cadence, where we assume a default period of every 2 epochs), emission values below a given brightness threshold (also a tunable hyperparameter, with a default value of 0.2), are clipped to 0. By assuming that very dim values are not light sources, we avoid local minima. At the same time, values that are above this threshold are likely to be light sources, and therefore, we boost these values, in order to accelerate the optimization, since extreme emission values in the 32-bit linear color space may be arbitrarily larger than 1. This part of the optimization should be self-correcting: values that are boosted when they shouldn't be will then be brought down in order to satisfy the rendering loss objective. See Figure 10 for an example of the converged result of optimizing the emission texture map. Finally, once the optimization has converged, we post process the mesh by pruning off any triangular faces whose vertices all fall below a threshold, by looking up the brightness of each vertex in the emission UV texture map, according to that vertex's associated UV coordinates. This results in a greatly reduced mesh that only contains faces that should emit light onto the rest of the scene, and can therefore be used as an area light for computing the shadow pass of our Hybrid Renderer.

## E. HDR Rendering Experiment Details

### E.1. Data acquisition

To train the HDR NeRF, we first construct a set of 32-bit HDR images of a scene, where each image is analogous to a single LDR image in the baseline NeRF formulation. To create a single one of these HDR images, we shoot a bracketed series of LDR exposures that consists of several images (7 in our experiments) captured from a fixed camera pose. We use a single Canon 5D MKIII with a 28mm lens, and set the camera to record images in 8-bit JPEG format. Specifically, the camera is mounted to a tripod to minimize any minute difference in camera pose between each of the exposures in the bracket, and to minimize any optical differences between the images (e.g. depth of field, bloom, etc.) apart from raw exposure value, we only adjust the exposure duration (i.e. shutter speed) between images, while the aperture remains fixed. The *relative* difference in exposure between any consecutive images in the bracket is kept constant, though the magnitude of this difference depends

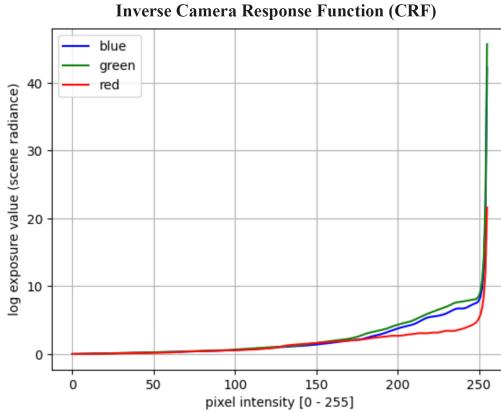


Figure 11: **Inverse Camera Response Function (CRF)** showing the recovered nonlinear function mapping from sRGB 8-bit pixel color to linear 32-bit radiance.

on the degrees of dynamic range of the scene being captured; ideally, within the range of exposures from darkest to brightest, every pixel of the image should be “well-exposed” in at least one of the images. We generally captured 7 images in the bracketed exposure, with 1-2 stops difference between each image.

## E.2. Data preprocessing

After capturing the scene, we preprocess the data using OpenCV. For each bracketed exposure consisting of  $n$  images taken from the same camera pose with different exposures, we align the images to ensure the best possible quality of the HDR image result, as even slight perturbations in the camera pose may cause degradation. Then, we select a single bracketed series that is “representative” of the lighting conditions of the full set, roughly covering the upper and lower limits of the overall dynamic range. The representative bracketed series is used to recover the camera’s nonlinear tone-mapping function, which we will refer to as the Camera Response Function (CRF) [Figure 11] using the method from [17].

To improve computational efficiency, we recover the CRF from only one representative image, although the linear system that must be solved could theoretically incorporate pixels from *all* of the bracketed series as constraints. By applying the recovered CRF uniformly to all series, we achieve consistent levels across all images. After applying the CRF to all bracketed series, we downsample each resulting 32-bit image by a factor of  $4x$ . Finally, for numerical stability during downstream training, we normalize all of the images by linearly scaling radiance values to the range of  $[0, 255]$ . While this may seem similar to reducing the HDR result back to LDR, all values are scaled uniformly so the *relative* differences between values are preserved, and they are **not** quantized to integer values. The resulting HDR images are then stored as EXR files.

Note that for performing camera pose estimation using COLMAP, as is widely done in the NeRF literature, we only use a single 8-bit LDR image from each bracketed series. Similarly, when training an LDR model for comparison against its HDR counterpart, we use a single image from each bracketed series, where all of the singleton images have the same exposure duration.

## E.3. HDR NeRF Implementation Details

Training the HDR NeRF is largely the same as training the baseline (LDR) NeRF. In fact, the Instant-NGP implementation already supports training with 32-bit HDR images in EXR format. We simply need to toggle flags ‘is\_hdr’ in Instant-NGP’s data loader and ‘linear\_colors’ in the GUI. Therefore, training is supervised purely in (scaled) linear color space, and correspondingly, the radiance component of the neural field is defined in linear color space. Optionally, during rendering only, we may choose to apply a tonemapping function uniformly to all of the accumulated pixel radiance values to obtain the final image in sRGB color space.

## F. Details about the Rendering comparison

We conducted a user study where participants were asked to rank the realism, quality, and correctness of rendered results from three methods, including ours, on the following examples: Mirror (main paper Fig. 2), Garden, and Bicycle (main paper

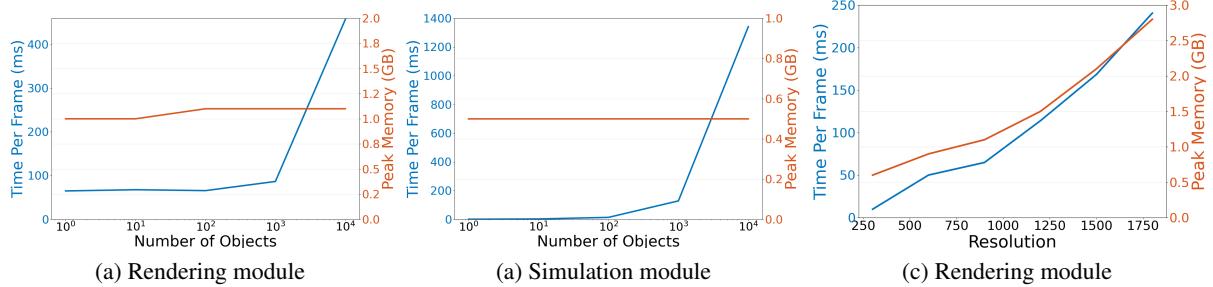


Figure 12: **Scaling the number of the inserted objects and image resolution.** When we increase the number of simulated objects from 1 to  $10^4$ , the peak simulation memory and rendering memory stay nearly constant. The (a) simulation time scales linearly w.r.t. the number of objects, while the (b) rendering time can increase when the number of the inserted meshes is too high. The (c) rendering time and memory usage in rendering modules scales linearly w.r.t. the number of pixels.



Figure 13: **Run time comparison with NeuPhysics [65].** We simulate a ball that bounces around (left). Compared to NeuPhysics, our method supports hybrid NeRF and mesh rendering and runs 25x-50x faster.

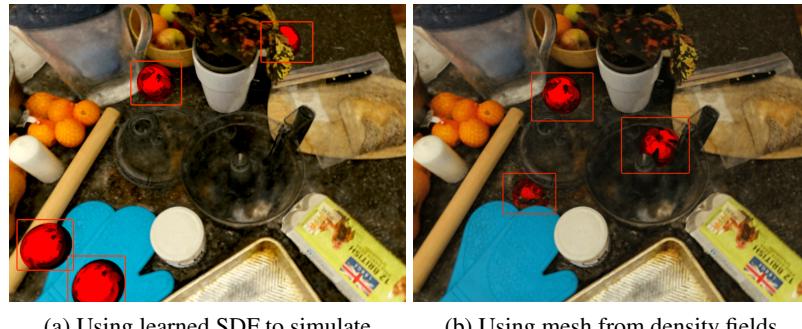


Figure 14: **Simulation Comparison.** Our simulation (a) using learned SDF as collision proxy geometry can correctly handle the contact between the counter and four balls. (b) Using marching-cube mesh directly from the density field, it fails to resolve collision and two of the four balls sink into the counter.

Fig. 5). Of 15 participants, 13/15 ranked our results highest for Mirror, while 12/15 did so for Garden and 15/15 for Bicycle (see Fig. 15). Binomial tests indicate that our images are statistically superior in all three scenes, with significant results at a significance level of 0.05. The corresponding p-values were found to be 0.035, 0.007, and  $6 \times 10^{-5}$ , respectively. We will include the quantitative results in the revised paper.

We also consider metrics such as FID and CLIP [5] scores. However, FID requires a ground truth distribution we cannot access. CLIP is a large language/vision model, but its scores are sensitive to prompts, and it is likely prone to misinterpreting reflections. Defining a scalable and robust quantitative metric to measure the visual fidelity of generative tasks remains an open problem.

MirrorNeRF designs a convenient capture system utilizing mirrors, but it does not focus on the rendering problem, thus does not overlap with the scope of our work.

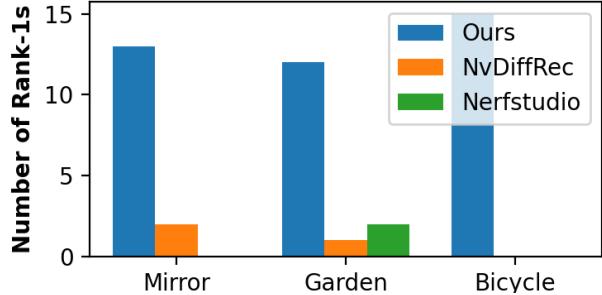


Figure 15: Number of Rank-1s each method receives.



Figure 16: Comparison between Luma AI (left) and Ours (right).

LumaAI recently released a [UE plugin](#) in April *after* the submission deadline. They provide binaries but no source code, so we only have a limited understanding of how it works. It is stated on their blog that NeRF cannot be used to simulate (otherwise, users need to add collider objects *manually*). Moreover, when we put a mirror in the Garden, shadows in the mirror reflections look strange. By contrast, our reflections and shadows are superior (see Fig. 16).