Eurographics Symposium on Rendering 2026
I. Gkioulekas and A. Jarabo
(Guest Editors)

COMPUTER GRAPHICS *forum*
Volume 45 (2026), Number 4

Lighting-Consistent Object Transfer Across Radiance Fields

N. Violante[1], G. Kopanas[2], L. Franke[1], J. Philip[3] and G. Drettakis[1]

[1]Inria, Université Côte d'Azur, France [2]Google DeepMind, United States [3]Eyeline Labs, United Kingdom

Figure 1: *We transfer an object from one captured 3DGS scene into another and harmonize its lighting so that shadows and reflections match the new environment. Naively compositing the source object (the car, from Ref-NeRF [VHM*21]) into the target scene (from DL3DV [LST*24]) leaves it inconsistently lit, with missing shadows and wrong reflection highlights (upper row). Our method harmonizes the object and produces a multi-view consistent result (lower row).*

Abstract

3D Gaussian Splatting (3DGS) is widely used to capture and render real scenes. Compositing objects from one capture into another has applications in many domains, such as VFX, architecture and interior design, or marketing. However, extracting an object from a source scene and naively pasting it into a target scene will fail to produce realistic results due to the different lighting conditions between the two scenes. To address this problem, we introduce a diffusion model that harmonizes naively composited images with inconsistent lighting. The model is trained with a heterogeneous dataset of image pairs (inconsistent composite input, consistent output), combining synthetic, generated, and real data. Our complete 3D solution allows a user to extract an object from the source scene and composite it into the target scene. From this, the (inconsistent) views of the target scene with the composite object are rendered. Our diffusion model harmonizes each one of these views, which are finally consolidated in a 3DGS representation with a post-optimization step. Our method provides visually compelling results, making object transfer between 3DGS easy to use and significantly improving quality compared to previous methods.

1. Introduction

3D Gaussian Splatting (3DGS) [KKLD23] allows easy capture and real-time free-viewpoint navigation for real scenes. Unfortunately, such representations are hard to edit, since they entangle materials and lighting as they appear at capture time. Compositing scenes and objects captured independently has many real-world applications ranging from VFX and architectural visualization to interior design and marketing. But naively compositing scenes by copying Gaussian primitives from a source scene to a target scene leads to unrealistic results because of the lighting inconsistencies between the two scenes (Fig. 1, top row). We introduce a method that allows a user to extract an object from one captured 3DGS scene, interactively place it into another scene in 3D, and produce a composited scene with consistent lighting.

Traditional solutions to the composition problem involve modeling the geometry and materials of the objects, and rendering using estimated light from the captured target scene; this is technically demanding and costly, requiring both expert 3D modeling and reliable lighting estimation. Previous automated methods have tried to harmonize the illumination of an inserted object for the 2D image case [NCL*25]. Recently, diffusion models have been used as priors to allow relighting of images, producing consistent highlights and shadows [JLL*24; CWP*25; MHT*25], and also to perform intrinsic image decomposition and editing [ZDG*24; KSN23; CMA23; CA25]. Lifting such 2D image methods to 3D requires making the result consistent over all input views used for capture. This is a hard task due to the inverse and ambiguous nature of the problem combined with the lack of hyper-realistic 3D datasets. Diffusion models can also relight captured 3D objects in the simplified case of distant illumination [JTL*24; LZF*24], but ignore cast shadows and foreground-background interactions. These models have been used to relight more general scenes [PGP*24], but composition and harmonization across captures are not feasible with such methods.

We thus present DOT3D, *Diffusion Object Transfer in 3D,* a method to transfer objects from a source scene into a target scene. DOT3D allows the user to interactively extract an object from a source 3DGS scene, place it at the desired location in a target 3DGS environment, and harmonize the object to match its new environment. To harmonize the object in the target scene, we first render each input view of the naively composited scene and use a fine-tuned diffusion model to harmonize the *inconsistent lighting* in these renders. We then consolidate the independently harmonized views in 3D so they are multi-view consistent, producing a fully harmonized 3DGS scene.

In practice, this process faces two challenges: 1) fine-tuning a diffusion model to harmonize images requires a high-quality dataset, and 2) ensuring multi-view consistency of the individually harmonized images. To address the first challenge, we create a heterogeneous dataset that combines three sources: synthetic data with ground-truth values rendered using global illumination, paired multi-illumination data generated by a high-quality diffusion model, and paired real data showing scenes with and without an object, augmented with a relighting network. For the second challenge, we propose a 3DGS post-optimization to consolidate the individually harmonized views into a consistent 3DGS representation.

In summary, our main contributions are:

- A diffusion model specialized for 2D lighting harmonization of image object insertion, trained with a heterogeneous dataset that includes multiple sources of synthetic, generated, and real data.
- A 3DGS post-optimization for multi-view consolidation of independently harmonized views to create a consistent scene.
- An interactive pipeline allowing users to extract an object from a source scene, and insert it into a target capture with consistent lighting.

Our results on a variety of commonly used datasets demonstrate that we achieve plausible and consistent object transfer between captured real scenes. Our code and data are available at `https://repo-sam.inria.fr/nerphys/dot3d`

2. Related Work

Novel View Synthesis with Radiance Fields. Neural Radiance Fields (NeRF) [MST*20; BMT*21; BMV*22] represent a scene with a multi-layer perceptron (MLP) that maps position and viewing direction into volumetric density and view-dependent color. The colors along pixel rays are integrated to produce the rendered image. Despite using acceleration structures [MESK22; FYT*22], rendering each pixel involves marching along a ray and querying the MLP hundreds of times, making rendering slow. To overcome this limitation, 3D Gaussian Splatting (3DGS) [KKLD23] represents a scene with a set of Gaussian primitives that can be efficiently rasterized on the GPU, allowing real-time rendering at over 100 FPS. Since its introduction, 3DGS has been extended to address some of its limitations, notably high-frequency reflections [YHZ24; KWT25], and has shown promising development for relighting and inverse rendering in the simplified object-centric case [JTL*24; LZF*24].

Radiance Field Segmentation. To extract an object from one scene and insert it into another, we must identify which Gaussians belong to the object. Most methods extend the Gaussian primitives with additional features, and then optimize these features to match 2D masks of the object across different views [SPB*23; QLZ*23; QYZW24; ZCJ*24; LYB*24; YDYK23; ZLL*25]. During optimization, contrastive methods check whether two pixels belong to the same mask. This avoids tracking mask correspondences across views, but makes optimization slow [GLF*24; KWK*24; CSK*24; CFY*24; YYZ*24; VMG*25]. Since we extract a single object, we optimize the features of the primitives to match a binary mask using a simple binary classification loss, making optimization faster.

Light Control with Generative Models. To insert an object into a new environment, we must relight the object and cast proper shadows and reflections to match the new environment. Early approaches based on GANs [KLA19; KLA*20] edit the lighting of generated scenes via latent space manipulation [HHLP20; BMHF23; WYLL22]. However, training GANs is prone to instabilities and mode collapse [BZW*19]. Diffusion models [SWMG15; HJA20; DN21; SE19; SSK*20] are more stable to train and offer unprecedented high-quality text-to-image synthesis at large scale [RBL*22; RDN*22]. These models require costly training, which has motivated fine-tuning and control methods that adapt pre-trained models to novel tasks [HSW*21; YZL*23; ZRA23].

Building on this progress, diffusion models have been widely adopted by the graphics and vision communities, and in particular, they have been extended for lighting manipulation tasks on images. Recent work has enabled direct control over light sources [PGP*24; MHT*25]. Another line of work, inspired by 3D modeling workflows, employs intrinsic decomposition methods [ZDG*24; KSN23; LCY*24] to extract channels such as albedo, roughness, normals, and shading, which in turn enable editing through direct manipulation of these channels. Applications include object insertion, removal, and relighting [LDH*25; ZFG*25; CMA23], but these approaches require precise edits on multiple channels, and are restricted to the 2D case.

Careaga and Aksoy [CA25] propose an intrinsic-based approach for 2D relighting with an intermediate mesh estimation to control lighting. Recently, initial efforts have been made to consolidate 2D

intrinsic decomposition into 3D [KHN25; LDL25], but with restrictive assumptions, e.g., complete mesh available, or isolated objects only lit by environment maps.

Composition and Inpainting. Composition methods aim to integrate objects given a coarse placement [JKK*23] but often struggle to preserve object identity and pose [SZL*22; YGZ*22; CHL*24]. Similar in spirit to our method, Nicolet et al. [NPD20] use a relighting network to harmonize lighting on images and rely on Unstructured Lumigraph [BBM*01] with meshes for novel view synthesis, thus restricting quality. ObjectDrop [WCF*24] enables precise placement and adds shadows and reflections to an inserted object. But unlike our work, it is restricted to 2D and does not account for the lighting of the inserted object, which generates inconsistent illumination when the object is taken from a scene with noticeable differences in lighting conditions.

Diffusion-based inpainting methods enable the generation of content within a user-specified region of an image [LDR*22]. To build a dataset with different lighting conditions, ControlCom [ZDL*23] applies only image-space operations (jitter, saturation, brightness) and is restricted to 2D. SpotLight's [FZM*24] data pipeline lacks generated data and real data, and the method needs guidance from shadow images.

In the 3D context, recent inpainting methods for 3DGS often focus on object removal [HCW25; LOW*24]. For object insertion, D3DR [SDF25] personalizes a diffusion model on a few images of the object [RLJ*23], but struggles to obtain a realistic integration of the object with its surroundings. MVInpainter [CYW*24] uses a multi-view network to inpaint multiple views. MV-CoLight [RBX*25] proposes two feed-forward transformers for harmonization: one for individual images, and another for Gaussian primitives. The method focuses on low-resolution images (256×256) and scenes with few images, typically between 6 and 16, covering limited points of view. In contrast, our method handles full scenes with hundreds of high-resolution images.

Diffusion for Novel-view and 3D Reconstruction. Both image and video diffusion models demonstrate outstanding generation capabilities. They are trained on massive datasets that have no equivalent in 3D, where data is scarce. Thus several works have explored the use of these pre-trained models to generate 3D content. CAT3D [GHH*24] leverages multi-view diffusion to generate a scene that is then baked into a 3DGS representation, while other approaches [RSH*25; YHXS25] use camera-conditioned video models. Complex high-level editing and generation can also be performed using the priors from diffusion models [HTE*23]. While designed for generative tasks, the natural image and video priors of these models have also been used to improve 3D reconstruction and novel view synthesis. The priors can help reconstruction in underconstrained regions with Gaussian artifacts [LZH24; WZT*25], in sparse view settings [WMH*24], and when content is missing [WXH*25]. Such models can also be fine-tuned to remove Gaussian-like artifacts as a post process, allowing very high resolution in constrained cases [PMC*25].

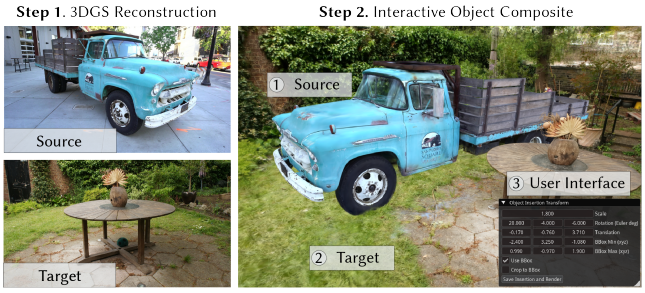

Figure 2: ***Object Composition Interface.*** *We propose a user interface for object composition across radiance fields. The user loads the extracted object from the source scene, inserts it into the target scene, and can control its final position, orientation, and scale within the target scene. Composite images are rendered using the camera viewpoints from the target scene.*

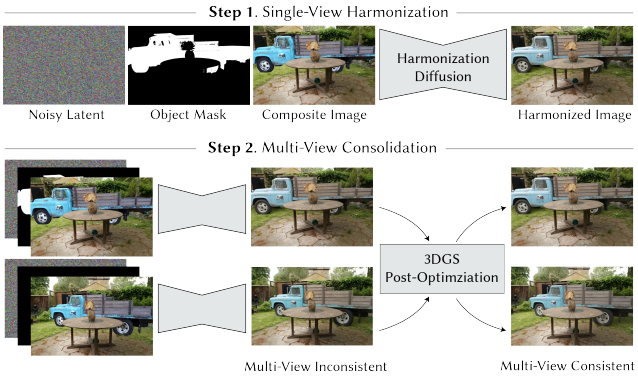

Figure 3: ***Method Overview.*** *Step 1: Single-View Harmonization: Conditioned on the composite image and the object mask, our harmonization model transforms a noisy latent into a harmonized image with plausible lighting, shadows, and reflections. Step 2: Multi-View Consolidation: Our 3DGS post-optimization consolidates the independent harmonized single-views into a multi-view consistent representation.*

3. Overview

Our method transfers an object from a source scene into a target scene, harmonizing it to match its new environment. An overview of our method is presented in Fig. 2 and Fig. 3.

We first reconstruct a 3DGS representation of the source scene and extract an object by lifting 2D masks to the Gaussian primitives, enabling 3D segmentation. We then use an interactive interface to insert the object in a target 3DGS scene, with full control of the object's position, orientation, and scale (Fig. 2).

To harmonize each view of the target scene with the composited object, our diffusion model takes as input a render of the naively composited scene together with a binary mask of the inserted object, and produces a harmonized output image. In this output image, the object matches the target scene's lighting and is visually integrated with its surroundings, including secondary effects such as shadows and reflections. To train our diffusion model, we con-

struct paired examples where the input shows the object under inconsistent lighting while the target output depicts the same object consistently inserted into the scene. This enables the model to learn how to transform lighting-inconsistent composites into harmonized images.

While this process yields harmonized images for individual views, the results are independent and thus not multi-view consistent. To address this, we propose a post-optimization of the composited 3DGS representation. Starting from this composited scene, we optimize the colors of the Gaussian primitives with a perceptual loss, while keeping all other Gaussian attributes (geometry, opacity) constant. This procedure enforces consistency while preserving fidelity, ultimately producing a consistent insertion of the object within the target scene.

4. Interactive Object Composition

Our approach begins by reconstructing the source scene and extracting the object of interest from it. To this end, we obtain 2D binary masks of the foreground object using a pre-trained BiRefNet [ZGF*24], although other options are also well-suited for this task (e.g., SAM models [KMR*23; KYD*23; RGH*24]). We extend 3DGS for 3D segmentation by assigning to each Gaussian primitive an extra feature, alongside the color attributes, of dimension $d = 1$, which can be efficiently rasterized on the GPU. To train these features, we rasterize them into per-pixel features and optimize a binary classification cross-entropy objective using the binary masks as targets.

We then segment the object in 3D from this source scene using a similarity threshold of 0.75 on the binary features (after sigmoid activation), and transfer it into a target scene using a 3D interface that we designed for this purpose (Fig. 2). This interface enables users to adjust the object's position and orientation within the target scene, and also provides tools to refine the segmentation through bounding box selection to remove unwanted artifacts. This produces an *inconsistent* composite of the object from the source scene into the target scene, which we later harmonize.

5. Single-View Harmonization

After compositing, the object is not yet properly integrated into the target scene due to the lack of consistent lighting. To address this, we propose a diffusion-based harmonization model to adapt the object's illumination and its surroundings to fit in the target scene (Fig. 3). Our harmonization model takes as input an image of the composite object under mismatched illumination, and outputs a harmonized image where the object's lighting matches that of the target scene and is integrated with shadows and reflections, yielding a result that appears naturally captured rather than naively composited. Fine-tuned diffusion models, which exploit the rich data priors of the base model, are a powerful way to handle problems such as ours, but a critical component for success is the curation of a dataset for the specific task.

To build our model, we follow DEGS [PMC*25] and fine-tune FLUX.1-schnell [BBB*25], a state-of-the-art rectified-flow model [LCB*22] for text-to-image generation. FLUX.1-schnell produces high-quality results in only 4 sampling steps, which makes it efficient to run on the many images of a single scene. For efficiency, the diffusion process operates in the latent space of a pre-trained variational autoencoder rather than in image space. The forward noising process interpolates the target latent with Gaussian noise, $\mathbf{z}_t = (1-t)\mathbf{z}_0 + t\epsilon$, where $\mathbf{z}_0$ is the latent of the harmonized (target) image, $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ is Gaussian noise, and t is the timestep.

The backward process is represented by a neural network $\mathbf{v}_\theta$ whose weights θ are learned during fine-tuning by minimizing a rectified flow-matching loss:

$$\mathcal{L}_\theta = \mathbb{E}\left[\|\mathbf{v}_\theta(\mathbf{z}_t, \mathbf{z}_{\text{comp}}, \mathbf{z}_{\text{mask}}, c, t) - (\epsilon - \mathbf{z}_0)\|_2^2\right] \quad (1)$$

The network regresses the velocity $\epsilon - \mathbf{z}_0$ and is conditioned on two extra inputs: the latent of the composite image $\mathbf{z}_{\text{comp}}$ and the latent of the object's mask $\mathbf{z}_{\text{mask}}$. These are concatenated channel-wise with the noisy latent $\mathbf{z}_t$ at the current timestep and, unlike $\mathbf{z}_t$, are not interpolated with noise. Following DEGS [PMC*25], we accommodate the extra inputs by repeating the first linear layer of the original network and copying its weights and biases to initialize the new layer. We fix the text conditioning c to the empty string. The same strategy has been used for UNet-based models [ZDG*24; MHT*25].

5.1. Data Preparation

To train our harmonization model, we need to form image pairs $(\mathbf{x}^{\text{comp}}, \mathbf{x}^{\text{out}})$, where the composite input image $\mathbf{x}^{\text{comp}}$ contains an inserted object with lighting that is inconsistent with the surrounding scene, and the ground truth target $\mathbf{x}^{\text{out}}$ shows the same object under correct scene illumination, including consistent lighting effects such as shadows and reflections.

The source object is identified in the source image with a binary mask $\mathbf{M}$. The inconsistent input image $\mathbf{x}^{\text{comp}}$ is created by rendering an auxiliary image $\mathbf{I}^{\text{incon}}$ where the object is rendered alone with different lighting. Then the object in image $\mathbf{I}^{\text{incon}}$ is pasted into the background image $\mathbf{x}^{\text{bg}}$ that has the original source lighting of the scene without the object (and thus without its shadows and reflections):

$$\mathbf{x}^{\text{comp}} = \mathbf{M} \odot \mathbf{I}^{\text{incon}} + (1 - \mathbf{M}) \odot \mathbf{x}^{\text{bg}} \quad (2)$$

By training on this image-to-image task, our model learns to map an inconsistent input image $\mathbf{x}^{\text{comp}}$, where the object appears under incorrect lighting, into a harmonized target image $\mathbf{x}^{\text{out}}$, where the object matches the illumination of the surrounding scene. Conditioning on the object mask $\mathbf{M}$ provides the object's extent to the model.

5.2. Datasets

Synthetic Data. We use 30 high-quality scenes of indoor environments including kitchens, living rooms, and offices constructed by 3D artists in Blender [Com18], and we keep two of them as a separate test set for our ablations. We manually annotate the scenes by placing cameras in suitable positions and selecting which objects will be used to produce composite images. For each camera, we define a small arc path from which 5 images are rendered. For a

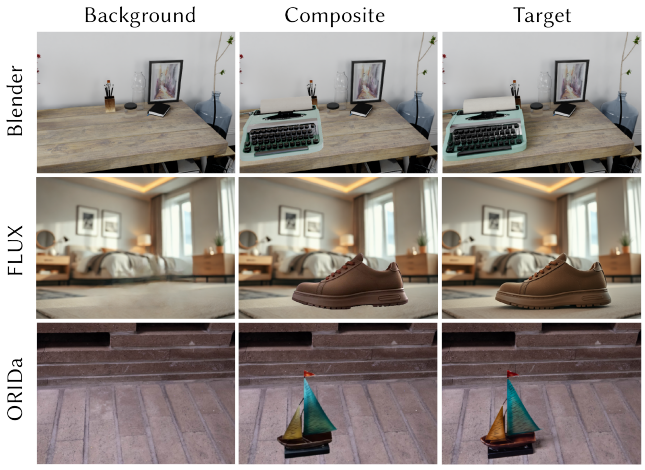

Figure 4: ***Data for Harmonization Model.*** *Following Eq. 2, we form an input image by compositing a background image of a scene with an object lit under a different illumination. The target image contains the harmonized object within the scene. During training, pairs of input and target images are sampled on-the-fly. We have three data sources: our synthetic dataset of Blender scenes, a dataset of images generated with FLUX [BBB*25], and the ORIDa dataset [KHJ*25].*

given camera and an object, our data generation pipeline first renders the image of the scene *with* the object, which we use as a consistent target. Next, we render the scene *without* the object to obtain a background. Finally, we sample an environment map and render the object in isolation under this different illumination. To form a composite image, we simply paste the isolated object into the background image using the object mask (Eq. 2). However, naively pasting the object could produce objects with large differences in exposure due to the different illumination conditions between the scene and the environment map. To address this, during training we normalize the image of the object under different illumination by the mean and standard deviation of the image of the object in the scene. We also apply this normalization to the composite images of the other two datasets.

All images are rendered in sRGB color space, using physically-based Monte Carlo path tracing (Blender Cycles [Com18]) with 128 spp at resolution 704×496. We automatically obtain object masks using the object ID from Blender. This dataset provides 1.9K consistent target images and over 51K inconsistent composite images from combinations of different objects and different environment maps. See Fig. 4 (top row) for an example.

Generated Data. We extend our dataset of synthetic images with generated images from the text-to-image model FLUX [BBB*25]. We use a set of text prompts to generate high-quality images of typical objects in different indoor environments. The text prompts combine different pre-defined categories of objects, environments, lighting conditions, and points of view to produce a variety of images. Unlike the synthetic dataset, in this case we do not have direct access to an image without the object to use as background, nor to object masks. To obtain the mask of the foreground object, we use BiRefNet [ZGF*24]. To address the lack of background images, we train a custom object removal network on our synthetic data. This network inputs an image of the scene with the object and a mask, and outputs a background image without the object. Most importantly, our object removal network also removes shadows and reflections caused by the object, which the harmonization model needs to learn to produce by itself. To produce the inconsistent versions of the object, we first relight the generated images using DiffusionRenderer [LGL*25] with different environment maps. Then we form the composite by segmenting the object from the relit image using the object mask, and pasting it on the background image. With this strategy, we generate 1.6K consistent target images, and 6.5K inconsistent composite images from the different relighting results. See Fig. 4 (middle row) for an example.

Real Data. To increase diversity in real data, we include captures from the object-centric ORIDa dataset [KHJ*25], which contains images of simple scenes with an object, its corresponding masks, and background images. However, the dataset is redundant: the exact same objects are used multiple times on the same backgrounds. For this reason, we subsample 2K images from the available 27K images to balance their contribution with respect to the other datasets. To produce the composites, we use DiffusionRenderer [LGL*25] to relight the images of the scene with the object. Then we produce the composite by segmenting the object from the relit image using the object mask, and pasting the object on the background image. This dataset provides 2K consistent target images, and 8K inconsistent composite images from the different relighting results. See Fig. 4 (bottom row) for an example.

5.3. Implementation Details

Harmonization Model Training. We fine-tune our harmonization model for $50K$ iterations on 4 NVIDIA H100 GPUs at resolution 704×496 using Adam [KB14] with a learning rate of 3×10^{-5} and a total batch size of 4, which takes about two days. We build our training and inference code with Diffusers [vPPL*22]. At inference, we use 4 denoising steps, which takes approximately 4 seconds on a single H100 for this resolution. Inference on new images is not restricted to the fine-tuning resolution (704×496). To handle arbitrary image resolutions, we pad them to make their resolution a multiple of 16 (required by the underlying FLUX architecture) and crop the result.

FLUX Dataset Generation. To generate the FLUX dataset (Sec. 5.2), we use the original FLUX.1-schnell [BBB*25] with 4 inference steps and a guidance scale factor of 1. To automate the process, we use the following text prompt:

"A realistic photo of a `object` in `environment` with `lighting`, `angle`, full view of the `object`, foreground object clearly shown, resting naturally on a surface or floor, with visible contact shadow, everything in focus."

The specific values of `object`, `environment`, `lighting`, and `angle` are taken from a set of predefined prompts.

To obtain the background images (without shadows and reflections) needed to create the (inconsistent) composites, we train a custom object removal network since generic models [BBB*25;

Rai25] prompted to remove an object would occasionally change other parts of the background. This network is fine-tuned from Stable Diffusion 2 [RBL*22] with the Diffusers codebase [vPPL*22] using our own synthetic dataset. We train the object removal network on a single NVIDIA H100 GPU with a batch size of 8 and a learning rate of 1×10^{-5} for 260K iterations.

6. Multi-View Consolidation

The input views of the target scene are independently harmonized after compositing the object. We run a post-optimization to consolidate those views into a multi-view consistent 3DGS representation (Fig. 3).

We initialize the representation from the composition of the object and the target scene, i.e., pasting the Gaussians of the object in the target scene. We optimize only the colors (all SH bands), keeping all other Gaussian attributes constant. Since the composited object is not necessarily seen from the same points of view in the source scene and in the target scene, we set the higher bands of SH to zero to reset the view-dependent effects. Otherwise, view-dependent effects from the source scene would remain from points of view that are not seen in the target scene. During post-optimization, we use a perceptual loss [ZIE*18] because its focus on higher-level semantic similarity, rather than pixel-wise photometric error, makes it more robust to potential inconsistencies across views [GHH*24].

Our implementation is based on the original 3DGS codebase [KKLD23], which includes the training speed acceleration from Taming-3DGS [MGK*24]. We run our post-optimization for 10K iterations with the default parameters of 3DGS. We also disable densification and opacity reset since the geometry is already consolidated.

In Section 7.1, we discuss and evaluate other consolidation strategies based on iterative dataset updates [HTE*23], a popular approach for radiance field editing, where the diffusion model is run multiple times on renderings from previously consolidated views, gradually converging to a multi-view consistent radiance field. Unlike some editing methods, which need to harmonize the color and the underlying geometry of the radiance field, we only need to harmonize its color. For this reason, it is easier for our post-optimization to rapidly converge to a multi-view consistent solution, without multiple updates of the dataset. The benefit of our post-optimization is that the harmonization model only needs to be run *once* per view.

7. Evaluation

We present qualitative evaluation on real scenes, quantitative evaluation using synthetic scenes where ground truth is available, and also ablations to better understand the importance of each component of our method.

Datasets. We demonstrate our method on a variety of scenes taken from Mip-NeRF-360 [BMV*22], Zip-NeRF [BMV*23], LERF [KKG*23], DL3DV [LST*24], Ref-NeRF [VHM*21], and Tanks and Temples [KPZK17]. We include some synthetic objects for insertion (plant, desk, locomotive) from BlenderKit, reconstructed with 3DGS. For quantitative results with a known ground truth, we provide nine extra high-quality synthetic scenes (see Fig. 7). We render images of these scenes with Blender Cycles [Com18], which are then used by the 3DGS reconstruction pipeline.

Qualitative 2D Evaluation. For image harmonization, we compare our harmonization model with LBM [CTSA25], MV-CoLight [RBX*25], and Nano Banana [Rai25] in Fig. 5. This allows us to assess both the overall quality of the generated images and the model's ability to generalize across different scenarios while maintaining consistent object integration and illumination adaptation. The limitations of these methods motivate training our own harmonization model.

Nano Banana [Rai25] sometimes exaggerates the shadows (e.g., first row, where the shadow of the truck is too strong for the scene), has color bleeding (fourth row, where the base of the pot has a green color bleeding effect) and is overall missing relighting and directional cues for providing realistic harmonization (e.g., third row, where the locomotive is not properly harmonized). To use Nano Banana as a harmonization model, we provide a prompt consisting of the composite image, the object mask, and the following text:

"I copy-pasted an object from one photograph to the other. I will give you the photograph and the mask of the copied object, white is the object and black is the background. Edit the photograph such that the copied object looks realistically composited in the environment. Create realistic contact shadows, ambient occlusion, global illumination, accurate lighting match, seamless blending etc. This is crucial: Do not change anything else other than the composited object."

LBM [CTSA25] is able to modify the object's lighting only for some images (truck, locomotive), but in all cases fails to produce realistic harmonization due to lack of shadows.

MV-CoLight [RBX*25] only produces satisfactory results for its in-distribution scenes from the DTC-MultiLight dataset (last two rows of Fig. 5), but fails to correctly harmonize out-of-distribution images of general real scenes.

Qualitative 3D Evaluation. To our knowledge, no prior method specifically addresses cross-scene 3DGS object transfer with lighting harmonization. Inverse rendering methods, which decompose scenes into intrinsic components to enable relighting, are therefore the most directly comparable approaches for this task, and we compare our method against Gaussian Shader [JTL*24], GS-IR [LZF*24], and 3DGS-DR [YHZ24], which we have extended to segment and insert an object from a source to a target scene. For other related methods, code is not publicly available ([RZX*25; ZYW*24]).

We show different insertion results in Fig. 6 on real scenes and Fig. 7 on synthetic scenes, demonstrating that our method effectively learns to harmonize the illumination of the object with its surroundings while seamlessly integrating it into the scene with consistent shadows and reflections. We evaluate our model on a wide range of scenes (both indoors and outdoors) and objects, covering diverse geometries, materials, and lighting conditions.

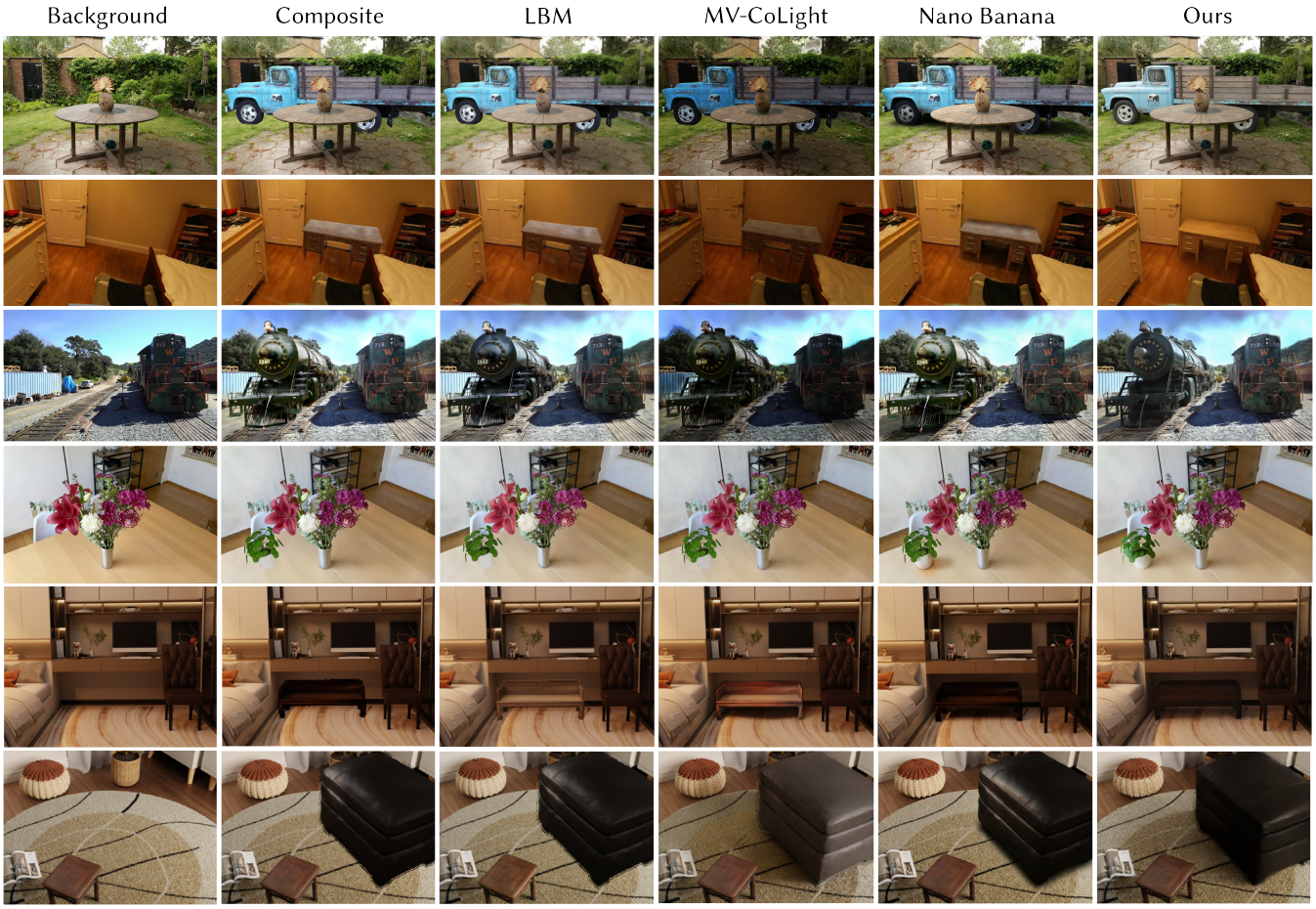

Figure 5: ***Qualitative 2D Results.*** *From left to right: Background, Composite, LBM [CTSA25], MV-CoLight [RBX*25], Nano Banana [Rai25], and Ours. From top to bottom, the composited scenes are: (i) Truck from Tanks and Temples into Garden from Mip-NeRF-360, (ii) Desk from BlenderKit into London Bedroom from Zip-NeRF, (iii) Locomotive from BlenderKit into Train from Tanks and Temples, (iv) Plant from BlenderKit into Bouquet from LERF. The last two scenes, (v) and (vi), are taken from the DTC-MultiLight dataset (MV-CoLight) [RBX*25].*

GS-IR [LZF*24] tends to produce overly saturated results on the object's surface, especially for white objects, as shown in Fig. 7 (e.g., the bed in the third row). Also, it struggles to adapt the overall lighting of the object to match the target scene, and it lacks contact shadows.

Gaussian Shader [JTL*24] produces more natural results without saturation on white objects but, similar to GS-IR, it does not adapt the lighting of the object and lacks shadows, resulting in unrealistic scenes.

3DGS-DR [YHZ24] struggles with the same issue: lack of proper harmonization and shadows, making the final result look unrealistic.

Overall, previous methods fail to correctly capture shadows and illumination conditions. In contrast, our method produces realistic object harmonization and creates proper shadows, making the insertion look natural. Fig. 7 contains cases with pronounced lighting mismatches: the composite sofa has a strong uneven illumination which our model correctly harmonizes; the bed has a strong highlight from the source scene which is also correctly handled. In Fig. 6, our method corrects the lighting in the front of the locomotive (fifth row), which is not achieved by other methods.

Quantitative 3D Evaluation. We evaluate our method on nine high-quality synthetic scenes: six indoors and three outdoors (Fig. 7). In Table 1, we report PSNR, SSIM, and LPIPS [ZIE*18] to measure reconstruction quality with respect to the ground truth. Since our harmonization model is at its core a generative model, it produces a *plausible* harmonization, which does not necessarily match the ground truth at a pixel level. To measure how similar the harmonized images are with respect to the ground truth at a distribution level, we also compute FID [HRU*17] and KID [BSAG18], which are typically used to evaluate the realism of generated images. Our method outperforms previous alternatives by a wide margin across all metrics.

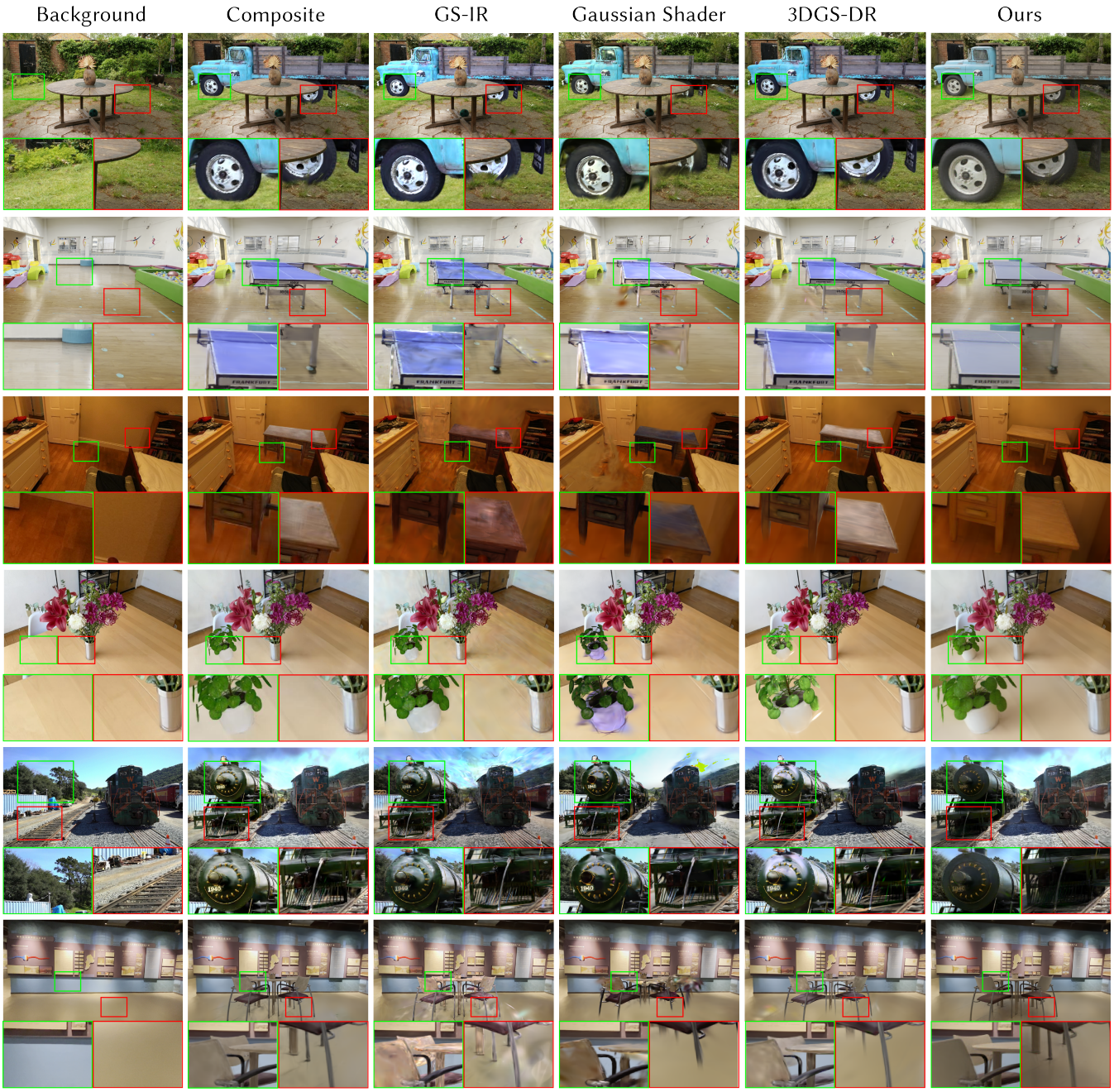

Figure 6: ***Qualitative 3D Results.*** *From left to right: Background, Composite, GS-IR [LZF*24], Gaussian Shader [JTL*24], 3DGS-DR [YHZ24], and Ours. From top to bottom, the composited scenes are: (i) Truck from Tanks and Temples into Garden from Mip-NeRF-360, (ii) 508850 into 4cea29, both from DL3DV, (iii) Desk from BlenderKit into London Bedroom from Zip-NeRF, (iv) Plant from BlenderKit into Bouquet from LERF, (v) Locomotive from BlenderKit into Train from Tanks and Temples, and (vi) d11d95 into 83d5f2, both from DL3DV. See the zoomed-in regions (green and red) for more details.*

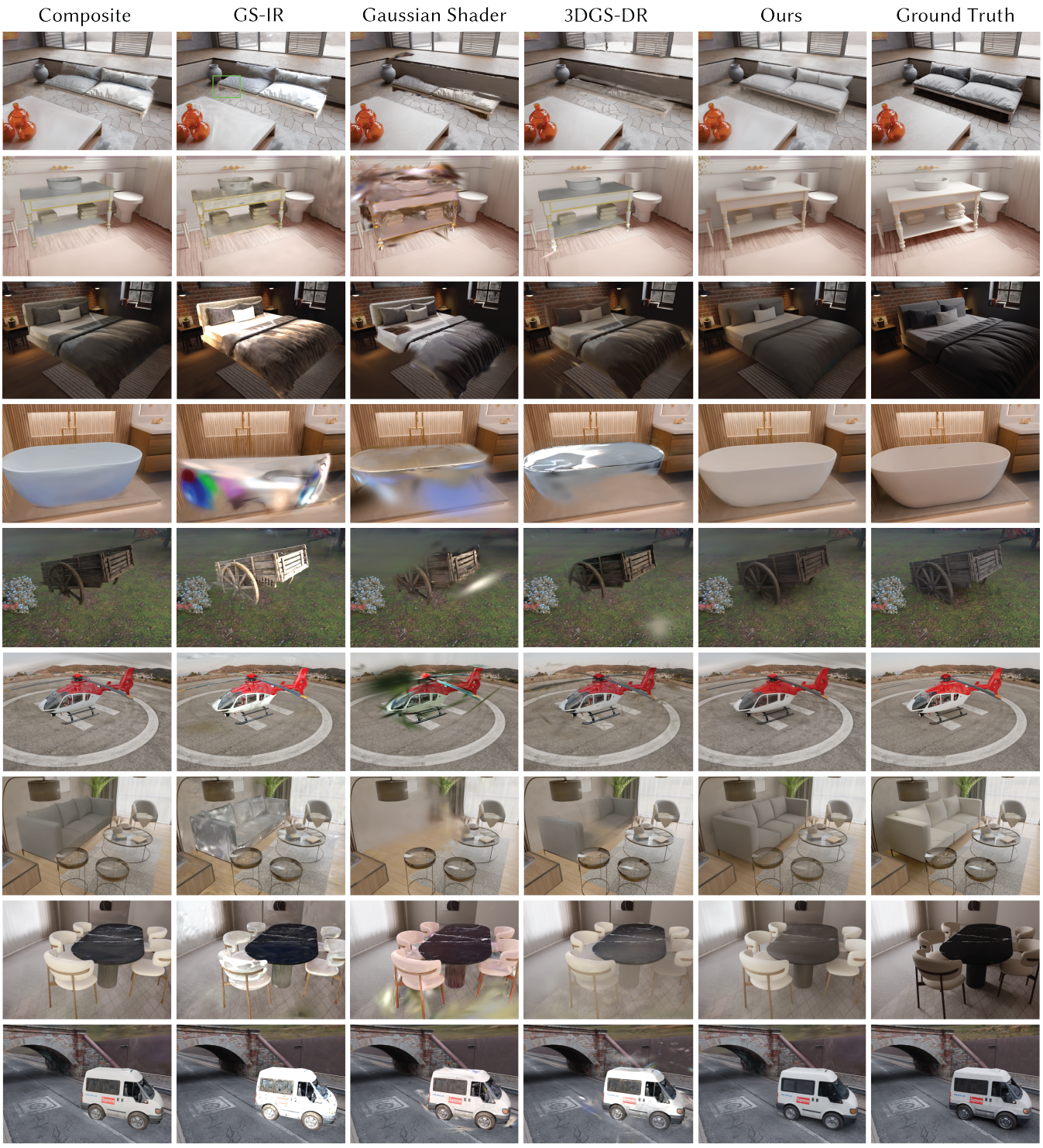

Figure 7: ***Qualitative 3D Results.*** *We provide nine high-quality synthetic scenes with ground truth for numerical comparisons. From left to right: Composite, GS-IR [LZF*24], Gaussian Shader [JTL*24], 3DGS-DR [YHZ24], Ours, and Ground Truth.*

Table 1: ***Quantitative 3D Evaluation.*** *We report metrics on nine synthetic scenes (Fig. 7), comparing our method against three inverse rendering techniques extended for object transfer across scenes: Gaussian Shader [JTL*24], GS-IR [LZF*24], and 3DGS-DR [YHZ24].*

	PSNR↑	SSIM↑	LPIPS↓	FID↓	KID↓
Ours	22.53	0.840	0.214	55.86	0.0413
GShader	18.66	0.799	0.272	152.62	0.1742
GS-IR	17.83	0.803	0.247	110.96	0.1143
3DGS-DR	19.65	0.814	0.245	124.07	0.1448

Table 2: ***Illumination Consistency.*** *We evaluate the illumination consistency of three scenes where an object has been inserted and rotated in-place in three different positions: P_0, P_1, P_2 (see Fig. 8). We report average results across the scenes for each position.*

	PSNR↑	SSIM↑	LPIPS↓	FID↓	KID↓
P_0	25.14	0.9135	0.1553	46.41	0.0161
P_1	24.85	0.9117	0.1589	38.34	0.0118
P_2	24.16	0.9097	0.1639	45.78	0.0160

Illumination Consistency. To evaluate the consistency of our illumination harmonization across different object configurations, we conduct an experiment on three synthetic scenes where we insert an object into each scene and rotate it across three poses (P_0, P_1, P_2). We run our harmonization model independently on each composite scene and then we run the post-optimization and compare with the GT for each position (we hold out every 8th image as a test image, following Mip-NeRF-360 [BMV*22]), obtaining similar metrics across scenes (Table 2). By comparing the harmonization and post-optimization with the ground truth (which is by definition multi-view consistent) in novel views, we are measuring the multi-view consistency of our approach. We show qualitative results in Fig. 8, where we observe a consistent illumination for the different positions of the objects in the scene. While a strong cast shadow is not produced for the bathroom, our model generates a plausible harmonization across views in all cases, including soft shadows on the walls.

7.1. Ablations

Data for Harmonization Model. We assess the impact of each one of our three data sources (Blender, FLUX, ORIDa) by training a harmonization model without each one of them, and evaluating the resulting model on a separate test split of each dataset. Our results, summarized in Table 3, show that our carefully curated Blender dataset has the most impact, significantly improving metrics. The FLUX dataset also has an important but lesser impact. The ORIDa dataset has the least impact overall: including it in the training process improves PSNR and FID slightly but worsens SSIM and LPIPS. We believe this is due to the limited diversity of the dataset, which consists of a short list of simple objects and backgrounds.

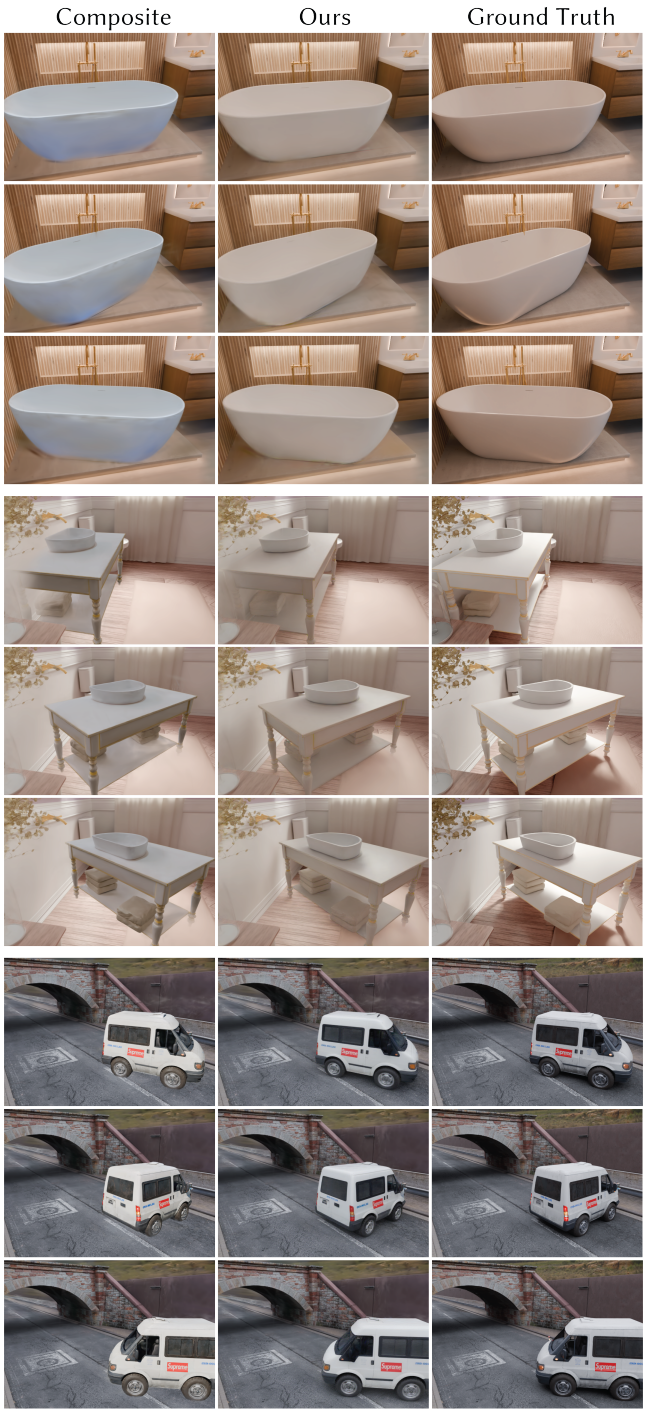

Figure 8: ***Illumination Consistency.*** *Results of our harmonization method on three synthetic scenes with an inserted object at three different rotations. Our method produces consistent multi-view illumination across all configurations.*

Table 3: ***Dataset Ablation.*** *We analyze the impact on our diffusion model of each one of the three data sources: synthetic data from Blender, generated images with FLUX [BBB*25], and real images from ORIDa [KHJ*25]. For each dataset, metrics are computed in a separate test set not seen during training.*

	Blender				FLUX				ORIDa				Average			
	PSNR↑	SSIM↑	LPIPS↓	FID↓	PSNR↑	SSIM↑	LPIPS↓	FID↓	PSNR↑	SSIM↑	LPIPS↓	FID↓	PSNR↑	SSIM↑	LPIPS↓	FID↓
Ours	27.08	0.893	0.124	33.86	28.34	0.892	0.108	25.13	30.08	0.865	0.170	74.64	28.50	0.883	0.134	44.54
w/o Blender	25.69	0.870	0.164	47.78	28.28	0.890	0.106	26.90	29.71	0.863	0.160	86.49	27.89	0.874	0.143	53.72
w/o Flux	26.98	0.895	0.122	34.64	24.70	0.852	0.147	46.23	30.41	0.868	0.163	70.19	27.36	0.872	0.144	50.36
w/o Orida	26.93	0.890	0.130	35.93	28.29	0.892	0.104	24.66	29.92	0.887	0.126	73.46	28.38	0.890	0.120	44.68

Table 4: ***Iterative Dataset Update Ablation.*** *We analyze the impact of using iterative dataset updates [HTE*23] with a linear and a geometric noise schedule with five dataset updates.*

	Post-opt ↓	PSNR↑	SSIM↑	LPIPS↓	FID↓	KID↓
Ours	15m30s	22.53	0.840	0.214	55.86	0.0413
w/Linear	1h8m	21.79	0.826	0.225	59.46	0.0462
w/Geometric	1h8m	21.55	0.820	0.230	60.47	0.0477

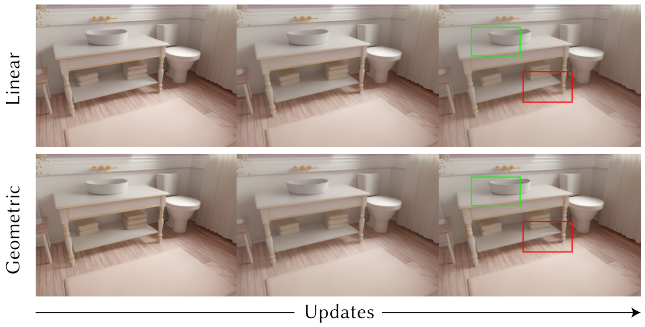

Figure 9: ***Iterative Dataset Update Ablation.*** *We experimented with two iterative dataset update schedules for the consolidation: linear from* 1.0 *to* 0.1 *and geometric from* 1.0 *to* 0.1. *In this example we perform five dataset updates (only updates 1, 3, and 5 are shown here). But we still found that a single consolidation produced satisfactory results without having to update the dataset multiple times (see Fig. 7).*

Iterative Dataset Update. We show that iterative dataset updates [HTE*23] are not necessary for object transfer across radiance fields: a simpler approach with a single update suffices to obtain realistic transfer results. An iterative approach alternates two steps: 1) applying the harmonization model to renders of already consolidated views (with decreasing values of starting noise), and 2) consolidating the current dataset of individually harmonized views. The dataset of views of the target scene is thus gradually updated, and the 3DGS representation converges to a multi-view consistent representation.

For all noise levels we use the composite image as conditioning for the harmonization model; early experiments using the subsequent renders as conditioning produced blurry results. We also try two different dataset update schedules: a linear and a geometric progression. In both cases, the first dataset update is done running our harmonization model starting from complete noise. But subsequent updates start from *lower* levels of noise by encoding the rendering from a given point of view in latent space and adding noise to it (e.g., adding noise at 80%). The diffusion process starts from this level of noise and proceeds until a clean image is obtained. Later updates use lower levels of noise, thus gradually converging to a multi-view consistent result.

We show a qualitative example (Fig. 9) comparing the geometric and the linear progression. But we opt for the more efficient solution of doing only one consolidation step, since it produces better results without having to update the dataset multiple times, thus being faster, as shown in Table 4.

Computational Cost. We run the single-view harmonization and the multi-view consolidation on a single NVIDIA H100 GPU. On our synthetic scenes of resolution 704×496 consisting of 200 images, the reconstruction of the source and target scenes takes 2m30s for each one. Training the binary features for segmentation for 10K iterations takes 40s. For the multi-view consolidation, running our harmonization model on the 200 composite images takes 12m30s, and the 3DGS post-optimization takes 3m.

Perceptual vs Photometric Loss. We use a perceptual loss [ZIE*18] for the 3DGS post-optimization. We observed improved results when compared to photometric error (L1/SSIM), which creates some small artifacts and more blurry results (qualitative examples in Fig. 10). On the other hand, post-optimization using a perceptual loss is slower than using a photometric loss. On our synthetic scenes of resolution 704×496 consisting of 200 images, using a single NVIDIA H100 GPU, the post-optimization takes 1m with L1/SSIM and 3m with the perceptual loss.

7.2. Limitations & Future Work

Our method is not without limitations. The variational autoencoder inherited from FLUX introduces some artifacts for objects with very high-frequency details, where the harmonized image has blur or block-like artifacts. Also, for some objects our harmonization network changes the appearance too much, altering the material properties. We show examples of these cases in Fig. 11.

Our method has no guarantee of absolute accuracy in the harmonization, only plausible results. Further scaling of the synthetic dataset with more scenes as well as capturing high-quality real data on complex scenes are promising directions to improve generalization. Finally, addressing the multi-view consistency with recent video models [KTZ*24; WWA*25; NAA*25] constitutes another interesting avenue for future research.

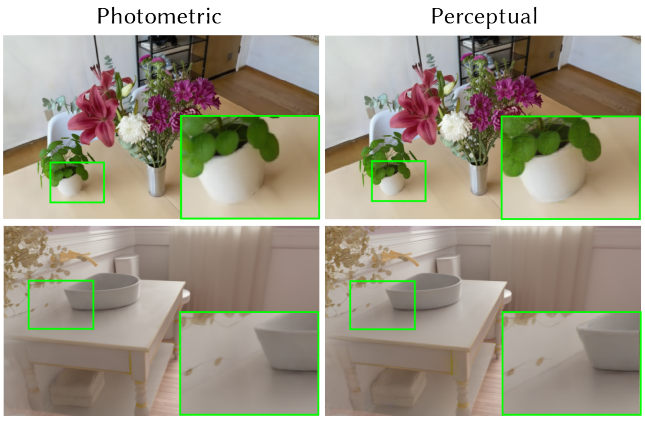

Figure 10: ***Photometric vs Perceptual Loss Ablation.*** *For 3DGS post-optimization, a perceptual loss better preserves details and sharp edges. See zoomed-in regions (green) for details.*

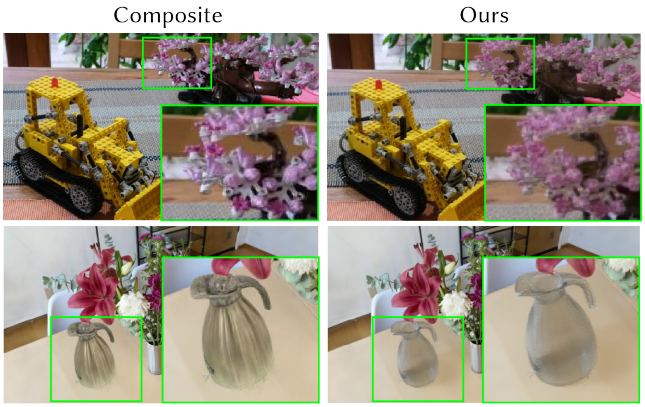

Figure 11: ***Limitations.*** *(top) Bonsai into Kitchen, both from Mip-NeRF-360 [BMV*22]. Contains block-like artifacts and blur due to very small details with high-frequency content. (bottom) Kettle from NERO [LWL*23] into Bouquet from LERF [KKG*23]. Despite the harmonization looking realistic, the appearance of the object changes from a metallic surface to glass.*

8. Conclusions

We presented DOT3D, a method that transfers 3D objects from one scene into another, maintaining consistent lighting including shadows and reflections. Our key contribution is a 3D lighting-consistent object transfer solution based on 3DGS, which first uses a fine-tuned diffusion model to harmonize the lighting of individual images after composition. These corrected views are consolidated in 3D via a post-optimization step, creating a full 3DGS scene with multi-view consistent lighting. We demonstrated our method on a variety of scenes from widely used datasets, and also provided quantitative results on high-quality synthetic scenes with ground truth. Our solution enables asset reuse across different scenes, ensuring easy and realistic integration in new environments, with applications spanning many fields.

Acknowledgments

This work was funded by the European Union, European Research Council (ERC) Advanced Grants NERPHYS, 101141721 https://project.inria.fr/nerphys and EXPLORER, 101097259 https://cordis.europa.eu/project/id/101097259. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them. Experiments presented in this paper were carried out using the Grid'5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several Universities as well as other organizations (see https://www.grid5000.fr). The authors would also like to thank Adobe and NVIDIA for software and hardware donations.

References

[BBB*25] Black Forest Labs, Batifol, Stephen, Blattmann, Andreas, et al. *FLUX.1 Kontext: Flow Matching for In-Context Image Generation and Editing in Latent Space*. 2025. arXiv: 2506.15742 [cs.GR]. URL: https://arxiv.org/abs/2506.15742 4, 5, 11.

[BBM*01] Buehler, Chris, Bosse, Michael, McMillan, Leonard, et al. "Unstructured Lumigraph Rendering". *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*. SIGGRAPH '01. New York, NY, USA: Association for Computing Machinery, 2001, 425–432. ISBN: 158113374X. DOI: 10.1145/383259.383309. URL: https://doi.org/10.1145/383259.383309 3.

[BMHF23] Bhattad, Anand, McKee, Daniel, Hoiem, Derek, and Forsyth, David. "Stylegan knows normal, depth, albedo, and more". *Advances in Neural Information Processing Systems* 36 (2023), 73082–73103 2.

[BMT*21] Barron, Jonathan T, Mildenhall, Ben, Tancik, Matthew, et al. "Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields". *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, 5855–5864 2.

[BMV*22] Barron, Jonathan T, Mildenhall, Ben, Verbin, Dor, et al. "Mip-nerf 360: Unbounded anti-aliased neural radiance fields". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, 5470–5479 2, 6, 10, 12.

[BMV*23] Barron, Jonathan T, Mildenhall, Ben, Verbin, Dor, et al. "Zip-nerf: Anti-aliased grid-based neural radiance fields". *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, 19697–19705 6.

[BSAG18] Bińkowski, Mikołaj, Sutherland, Danica J, Arbel, Michael, and Gretton, Arthur. "Demystifying mmd gans". *arXiv preprint arXiv:1801.01401* (2018) 7.

[BZW*19] Bau, David, Zhu, Jun-Yan, Wulff, Jonas, et al. "Seeing what a gan cannot generate". *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, 4502–4511 2.

[CA25] Careaga, Chris and Aksoy, Yağiz. "Physically Controllable Relighting of Photographs". *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*. SIGGRAPH Conference Papers '25. Association for Computing Machinery, 2025. ISBN: 9798400715402. DOI: 10.1145/3721238.3730666 2.

[CFY*24] Cen, Jiazhong, Fang, Jiemin, Yang, Chen, et al. *Segment Any 3D Gaussians*. arXiv:2312.00860. May 2024. DOI: 10.48550/arXiv.2312.00860. URL: http://arxiv.org/abs/2312.00860 (visited on 11/21/2024) 2.

[CHL*24] CHEN, XI, HUANG, LIANGHUA, LIU, YU, et al. *AnyDoor: Zero-shot Object-level Image Customization*. May 2024. DOI: 10.48550/arXiv.2307.09481. arXiv: 2307.09481 [cs]. (Visited on 09/25/2025) 3.

[CMA23] CAREAGA, CHRIS, MIANGOLEH, S. MAHDI H., and AKSOY, YAĞIZ. "Intrinsic Harmonization for Illumination-Aware Image Compositing". *SIGGRAPH Asia 2023 Conference Papers*. SA '23. Sydney, NSW, Australia: Association for Computing Machinery, 2023. ISBN: 9798400703157. DOI: 10.1145/3610548.3618178 2.

[Com18] COMMUNITY, BLENDER ONLINE. *Blender - a 3D modelling and rendering package*. Blender Foundation. Stichting Blender Foundation, Amsterdam, 2018. URL: http://www.blender.org 4–6.

[CSK*24] CHOI, SEOKHUN, SONG, HYEONSEOP, KIM, JAECHUL, et al. "Click-Gaussian: Interactive Segmentation to Any 3D Gaussians". *ECCV*. 2024 2.

[CTSA25] CHADEBEC, CLÉMENT, TASAR, ONUR, SREETHARAN, SANJEEV, and AUBIN, BENJAMIN. "LBM: Latent Bridge Matching for Fast Image-to-Image Translation". *arXiv preprint arXiv:2503.07535* (2025) 6, 7.

[CWP*25] CHOI, JUN MYEONG, WANG, ANNIE, PEERS, PIETER, et al. "ScribbleLight: Single Image Indoor Relighting with Scribbles". *IEEE International Conference on Computer Vision and Pattern Recognition*. June 2025 2.

[CYW*24] CAO, CHENJIE, YU, CHAOHUI, WANG, FAN, et al. *MVInpainter: Learning Multi-View Consistent Inpainting to Bridge 2D and 3D Editing*. Nov. 2024. DOI: 10.48550/arXiv.2408.08000. arXiv: 2408.08000 [cs]. (Visited on 09/27/2025) 3.

[DN21] DHARIWAL, PRAFULLA and NICHOL, ALEX. *Diffusion Models Beat GANs on Image Synthesis*. June 2021. DOI: 10.48550/arXiv.2105.05233. arXiv: 2105.05233 [cs, stat]. (Visited on 02/21/2024) 2.

[FYT*22] FRIDOVICH-KEIL, SARA, YU, ALEX, TANCIK, MATTHEW, et al. "Plenoxels: Radiance Fields without Neural Networks". *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. New Orleans, LA, USA: IEEE, June 2022, 5491–5500. ISBN: 978-1-6654-6946-3. DOI: 10.1109/CVPR52688.2022.00542. (Visited on 02/26/2024) 2.

[FZM*24] FORTIER-CHOUINARD, FRÉDÉRIC, ZHANG, ZITIAN, MESSIER, LOUIS-ETIENNE, et al. "Spotlight: Shadow-guided object relighting via diffusion". *arXiv preprint arXiv:2411.18665* (2024) 3.

[GHH*24] GAO, RUIQI, HOLYNSKI, ALEKSANDER, HENZLER, PHILIPP, et al. *CAT3D: Create Anything in 3D with Multi-View Diffusion Models*. May 2024. DOI: 10.48550/arXiv.2405.10314. arXiv: 2405.10314 [cs]. (Visited on 05/17/2024) 3, 6.

[GLF*24] GU, QIAO, LV, ZHAOYANG, FROST, DUNCAN, et al. *EgoLifter: Open-world 3D Segmentation for Egocentric Perception*. en. arXiv:2403.18118 [cs]. Mar. 2024. URL: http://arxiv.org/abs/2403.18118 (visited on 04/16/2024) 2.

[HCW25] HUANG, SHENG-YU, CHOU, ZI-TING, and WANG, YU-CHIANG FRANK. *3D Gaussian Inpainting with Depth-Guided Cross-View Consistency*. Apr. 2025. DOI: 10.48550/arXiv.2502.11801. arXiv: 2502.11801 [cs]. (Visited on 09/27/2025) 3.

[HHLP20] HÄRKÖNEN, ERIK, HERTZMANN, AARON, LEHTINEN, JAAKKO, and PARIS, SYLVAIN. "Ganspace: Discovering interpretable gan controls". *Advances in Neural Information Processing Systems* 33 (2020), 9841–9850 2.

[HJA20] HO, JONATHAN, JAIN, AJAY, and ABBEEL, PIETER. *Denoising Diffusion Probabilistic Models*. Dec. 2020. DOI: 10.48550/arXiv.2006.11239. arXiv: 2006.11239 [cs, stat]. (Visited on 02/21/2024) 2.

[HRU*17] HEUSEL, MARTIN, RAMSAUER, HUBERT, UNTERTHINER, THOMAS, et al. "Gans trained by a two time-scale update rule converge to a local nash equilibrium". *Advances in neural information processing systems* 30 (2017) 7.

[HSW*21] HU, EDWARD J., SHEN, YELONG, WALLIS, PHILLIP, et al. *LoRA: Low-Rank Adaptation of Large Language Models*. Oct. 2021. DOI: 10.48550/arXiv.2106.09685. arXiv: 2106.09685 [cs]. (Visited on 07/29/2024) 2.

[HTE*23] HAQUE, AYAAN, TANCIK, MATTHEW, EFROS, ALEXEI A, et al. "Instruct-nerf2nerf: Editing 3d scenes with instructions". *Proceedings of the IEEE/CVF international conference on computer vision*. 2023, 19740–19750 3, 6, 11.

[JKK*23] JAMBON, CLÉMENT, KERBL, BERNHARD, KOPANAS, GEORGIOS, et al. "NeRFshop: Interactive Editing of Neural Radiance Fields". *Proceedings of the ACM on Computer Graphics and Interactive Techniques* 6.1 (May 2023). URL: https://repo-sam.inria.fr/fungraph/nerfshop/ 3.

[JLL*24] JIN, HAIAN, LI, YUAN, LUAN, FUJUN, et al. "Neural Gaffer: Relighting Any Object via Diffusion". *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. Nov. 2024. (Visited on 11/11/2024) 2.

[JTL*24] JIANG, YINGWENQI, TU, JIADONG, LIU, YUAN, et al. "GaussianShader: 3D Gaussian Splatting with Shading Functions for Reflective Surfaces". *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Seattle, WA, USA: IEEE, June 2024, 5322–5332. ISBN: 979-8-3503-5300-6. DOI: 10.1109/CVPR52733.2024.00509. (Visited on 06/19/2025) 2, 6–10.

[KB14] KINGMA, DIEDERIK P and BA, JIMMY. "Adam: A method for stochastic optimization". *arXiv preprint arXiv:1412.6980* (2014) 5.

[KHJ*25] KIM, JINWOO, HAN, SANGMIN, JEONG, JINHO, et al. *ORIDa: Object-centric Real-world Image Composition Dataset*. June 2025. DOI: 10.48550/arXiv.2506.08964. arXiv: 2506.08964 [cs]. (Visited on 09/25/2025) 5, 11.

[KHN25] KOCSIS, PETER, HÖLLEIN, LUKAS, and NIESSNER, MATTHIAS. "Intrinsic Image Fusion for Multi-View 3D Material Reconstruction". *arXiv preprint arXiv:2512.13157* (2025) 3.

[KKG*23] KERR, JUSTIN, KIM, CHUNG MIN, GOLDBERG, KEN, et al. "LERF: Language Embedded Radiance Fields". en. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. Paris, France: IEEE, Oct. 2023, 19672–19682. ISBN: 9798350307184. DOI: 10.1109/ICCV51070.2023.01807. URL: https://ieeexplore.ieee.org/document/10376596/ (visited on 03/11/2024) 6, 12.

[KKLD23] KERBL, BERNHARD, KOPANAS, GEORGIOS, LEIMKÜHLER, THOMAS, and DRETTAKIS, GEORGE. "3D Gaussian Splatting for Real-Time Radiance Field Rendering". Vol. 42. 4. July 2023. URL: https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/ 1, 2, 6.

[KLA*20] KARRAS, TERO, LAINE, SAMULI, AITTALA, MIIKA, et al. "Analyzing and improving the image quality of stylegan". *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, 8110–8119 2.

[KLA19] KARRAS, TERO, LAINE, SAMULI, and AILA, TIMO. "A style-based generator architecture for generative adversarial networks". *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, 4401–4410 2.

[KMR*23] KIRILLOV, ALEXANDER, MINTUN, ERIC, RAVI, NIKHILA, et al. "Segment Anything". *arXiv:2304.02643* (2023) 4.

[KPZK17] KNAPITSCH, ARNO, PARK, JAESIK, ZHOU, QIAN-YI, and KOLTUN, VLADLEN. "Tanks and temples: benchmarking large-scale scene reconstruction". *ACM Trans. Graph.* 36.4 (July 2017). ISSN: 0730-0301. DOI: 10.1145/3072959.3073599. URL: https://doi.org/10.1145/3072959.3073599 6.

[KSN23] KOCSIS, PETER, SITZMANN, VINCENT, and NIESSNER, MATTHIAS. "Intrinsic Image Diffusion for Indoor Single-view Material Estimation". *arXiv preprint arXiv:2312.12274* (2023) 2.

[KTZ*24] KONG, WEIJIE, TIAN, QI, ZHANG, ZIJIAN, et al. "Hunyuanvideo: A systematic framework for large video generative models". *arXiv preprint arXiv:2412.03603* (2024) 11.

[KWK*24] KIM, CHUNG MIN, WU, MINGXUAN, KERR, JUSTIN, et al. *GARField: Group Anything with Radiance Fields*. arXiv:2401.09419 [cs]. Jan. 2024. DOI: 10.48550/arXiv.2401.09419. URL: http://arxiv.org/abs/2401.09419 (visited on 03/04/2024) 2.

[KWT25] KOUROS, GEORGIOS, WU, MINYE, and TUYTELAARS, TINNE. *RGS-DR: Reflective Gaussian Surfels with Deferred Rendering for Shiny Objects*. May 2025. DOI: 10.48550/arXiv.2504.18468. arXiv: 2504.18468 [cs]. (Visited on 06/20/2025) 2.

[KYD*23] KE, LEI, YE, MINGQIAO, DANELLJAN, MARTIN, et al. "Segment Anything in High Quality". *NeurIPS*. 2023 4.

[LCB*22] LIPMAN, YARON, CHEN, RICKY TQ, BEN-HAMU, HELI, et al. "Flow matching for generative modeling". *arXiv preprint arXiv:2210.02747* (2022) 4.

[LCY*24] LUO, JUNDAN, CEYLAN, DUYGU, YOON, JAE SHIN, et al. "IntrinsicDiffusion: Joint Intrinsic Layers from Latent Diffusion Models". *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers '24*. Denver CO USA: ACM, July 2024, 1–11. ISBN: 979-8-4007-0525-0. DOI: 10.1145/3641519.3657472. (Visited on 07/22/2024) 2.

[LDH*25] LYU, LINJIE, DESCHAINTRE, VALENTIN, HOLD-GEOFFROY, YANNICK, et al. "IntrinsicEdit: Precise Generative Image Manipulation in Intrinsic Space". *ACM Trans. Graph.* 44.4 (July 2025), 106:1–106:13. ISSN: 0730-0301. DOI: 10.1145/3731173. (Visited on 09/14/2025) 2.

[LDL25] LANGSTEINER, PHILIPP, DIHLMANN, JAN-NIKLAS, and LENSCH, HENDRIK. "MatSpray: Fusing 2D Material World Knowledge on 3D Geometry". *arXiv preprint arXiv:2512.18314* (2025) 3.

[LDR*22] LUGMAYR, ANDREAS, DANELLJAN, MARTIN, ROMERO, ANDRES, et al. "Repaint: Inpainting using denoising diffusion probabilistic models". *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, 11461–11471 3.

[LGL*25] LIANG, RUOFAN, GOJCIC, ZAN, LING, HUAN, et al. *DiffusionRenderer: Neural Inverse and Forward Rendering with Video Diffusion Models*. Jan. 2025. DOI: 10.48550/arXiv.2501.18590. arXiv: 2501.18590 [cs]. (Visited on 02/03/2025) 5.

[LOW*24] LIU, ZHIHENG, OUYANG, HAO, WANG, QIUYU, et al. *InFusion: Inpainting 3D Gaussians via Learning Depth Completion from Diffusion Prior*. Apr. 2024. DOI: 10.48550/arXiv.2404.11613. arXiv: 2404.11613 [cs]. (Visited on 09/27/2025) 3.

[LST*24] LING, LU, SHENG, YICHEN, TU, ZHI, et al. "Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, 22160–22169 1, 6.

[LWL*23] LIU, YUAN, WANG, PENG, LIN, CHENG, et al. "NeRO: Neural Geometry and BRDF Reconstruction of Reflective Objects from Multi-view Images". *ACM Transactions on Graphics* 42.4 (July 2023), 114:1–114:22. ISSN: 0730-0301. DOI: 10.1145/3592134. (Visited on 02/26/2024) 12.

[LYB*24] LEE, HYUNJEE, YUN, YOUNGSIK, BAE, JEONGMIN, et al. *Rethinking Open-Vocabulary Segmentation of Radiance Fields in 3D Space*. en. arXiv:2408.07416 [cs]. Aug. 2024. URL: http://arxiv.org/abs/2408.07416 (visited on 08/19/2024) 2.

[LZF*24] LIANG, ZHIHAO, ZHANG, QI, FENG, YING, et al. "GS-IR: 3D Gaussian Splatting for Inverse Rendering". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, 21644–21653. (Visited on 11/07/2025) 2, 6–10.

[LZH24] LIU, XI, ZHOU, CHAOYI, and HUANG, SIYU. *3DGS-Enhancer: Enhancing Unbounded 3D Gaussian Splatting with View-consistent 2D Diffusion Priors*. Oct. 2024. DOI: 10.48550/arXiv.2410.16266. arXiv: 2410.16266 [cs]. (Visited on 03/04/2025) 3.

[MESK22] MÜLLER, THOMAS, EVANS, ALEX, SCHIED, CHRISTOPH, and KELLER, ALEXANDER. "Instant neural graphics primitives with a multiresolution hash encoding". *arXiv preprint arXiv:2201.05989* (2022) 2.

[MGK*24] MALLICK, SASWAT SUBHAJYOTI, GOEL, RAHUL, KERBL, BERNHARD, et al. "Taming 3dgs: High-quality radiance fields with limited resources". *SIGGRAPH Asia 2024 Conference Papers*. 2024, 1–11 6.

[MHT*25] MAGAR, NADAV, HERTZ, AMIR, TABELLION, ERIC, et al. *LightLab: Controlling Light Sources in Images with Diffusion Models*. May 2025. DOI: 10.1145/3721238.3730696. arXiv: 2505.09608 [cs]. (Visited on 06/04/2025) 2, 4.

[MST*20] MILDENHALL, BEN, SRINIVASAN, PRATUL P, TANCIK, MATTHEW, et al. "Nerf: Representing scenes as neural radiance fields for view synthesis". *European conference on computer vision*. Springer. 2020, 405–421 2.

[NAA*25] NVIDIA, AGARWAL, NIKET, ALI, ARSLAN, et al. *Cosmos World Foundation Model Platform for Physical AI*. Jan. 2025. DOI: 10.48550/arXiv.2501.03575. arXiv: 2501.03575 [cs]. (Visited on 02/13/2025) 11.

[NCL*25] NIU, LI, CONG, WENYAN, LIU, LIU, et al. *Making Images Real Again: A Comprehensive Survey on Deep Image Composition*. 2025. arXiv: 2106.14490 [cs.CV]. URL: https://arxiv.org/abs/2106.14490 2.

[NPD20] NICOLET, BAPTISTE, PHILIP, JULIEN, and DRETTAKIS, GEORGE. "Repurposing a Relighting Network for Realistic Compositions of Captured Scenes". *Proceedings of the ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games*. ACM. May 2020. URL: http://www-sop.inria.fr/reves/Basilic/2020/NPD20a 3.

[PGP*24] POIRIER-GINTER, YOHAN, GAUTHIER, ALBAN, PHILIP, JULIEN, et al. "A Diffusion Approach to Radiance Field Relighting using Multi-Illumination Synthesis". *Computer Graphics Forum*. Vol. 43. 4. Wiley Online Library. 2024, e15147 2.

[PMC*25] PHILIP, JULIEN, MA, LI, CLAUSEN, PASCAL, et al. "Detail Enhanced Gaussian Splatting for Large-Scale Volumetric Capture". *ACM Trans. Graph.* 44.6 (Dec. 2025). ISSN: 0730-0301. DOI: 10.1145/3763336. URL: https://doi.org/10.1145/3763336 3, 4.

[QLZ*23] QIN, MINGHAN, LI, WANHUA, ZHOU, JIAWEI, et al. *LangSplat: 3D Language Gaussian Splatting*. arXiv:2312.16084 [cs]. Dec. 2023. DOI: 10.48550/arXiv.2312.16084. URL: http://arxiv.org/abs/2312.16084 (visited on 03/11/2024) 2.

[QYZW24] QIU, RI-ZHAO, YANG, GE, ZENG, WEIJIA, and WANG, XIAOLONG. "Language-Driven Physics-Based Scene Synthesis and Editing via Feature Splatting". *European Conference on Computer Vision (ECCV)*. 2024 2.

[Rai25] RAISINGHANI, NAINA. *Introducing Nano Banana Pro*. Accessed: 2026-01-19. Nov. 2025. URL: https://blog.google/innovation-and-ai/products/nano-banana-pro/ 6, 7.

[RBL*22] ROMBACH, ROBIN, BLATTMANN, ANDREAS, LORENZ, DOMINIK, et al. *High-Resolution Image Synthesis with Latent Diffusion Models*. Apr. 2022. DOI: 10.48550/arXiv.2112.10752. arXiv: 2112.10752 [cs]. (Visited on 02/21/2024) 2, 6.

[RBX*25] REN, KERUI, BAI, JIAYANG, XU, LINNING, et al. *MV-CoLight: Efficient Object Compositing with Consistent Lighting and Shadow Generation*. May 2025. DOI: 10.48550/arXiv.2505.21483. arXiv: 2505.21483 [cs]. (Visited on 11/18/2025) 3, 6, 7.

[RDN*22] RAMESH, ADITYA, DHARIWAL, PRAFULLA, NICHOL, ALEX, et al. "Hierarchical text-conditional image generation with clip latents". *arXiv preprint arXiv:2204.06125* 1.2 (2022), 3 2.

[RGH*24] RAVI, NIKHILA, GABEUR, VALENTIN, HU, YUAN-TING, et al. "SAM 2: Segment Anything in Images and Videos". *arXiv preprint arXiv:2408.00714* (2024). URL: https://arxiv.org/abs/2408.00714 4.

[RLJ*23] RUIZ, NATANIEL, LI, YUANZHEN, JAMPANI, VARUN, et al. "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation". *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2023, 22500–22510 3.

[RSH*25] REN, XUANCHI, SHEN, TIANCHANG, HUANG, JIAHUI, et al. *GEN3C: 3D-Informed World-Consistent Video Generation with Precise Camera Control*. Mar. 2025. DOI: 10.48550/arXiv.2503.03751. arXiv: 2503.03751 [cs]. (Visited on 03/06/2025) 3.

[RZX*25] REN, CHENGWEI, ZHANG, FAN, XU, LIANGCHAO, et al. "GauUpdate: New Object Insertion in 3D Gaussian Fields with Consistent Global Illumination". *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2025, 28653–28663. (Visited on 11/14/2025) 6.

[SDF25] SKOROKHODOV, VSEVOLOD, DURASOV, NIKITA, and FUA, PASCAL. *D3DR: Lighting-Aware Object Insertion in Gaussian Splatting*. Mar. 2025. DOI: 10.48550/arXiv.2503.06740. arXiv: 2503.06740 [cs]. (Visited on 09/27/2025) 3.

[SE19] SONG, YANG and ERMON, STEFANO. "Generative modeling by estimating gradients of the data distribution". *Advances in neural information processing systems* 32 (2019) 2.

[SPB*23] SIDDIQUI, YAWAR, PORZI, LORENZO, BULÒ, SAMUEL ROTA, et al. "Panoptic Lifting for 3D Scene Understanding with Neural Fields". en. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Vancouver, BC, Canada: IEEE, June 2023, 9043–9052. ISBN: 9798350301298. DOI: 10.1109/CVPR52729.2023.00873. URL: https://ieeexplore.ieee.org/document/10203291/ (visited on 03/22/2024) 2.

[SSK*20] SONG, YANG, SOHL-DICKSTEIN, JASCHA, KINGMA, DIEDERIK P, et al. "Score-based generative modeling through stochastic differential equations". *arXiv preprint arXiv:2011.13456* (2020) 2.

[SWMG15] SOHL-DICKSTEIN, JASCHA, WEISS, ERIC, MAHESWARANATHAN, NIRU, and GANGULI, SURYA. "Deep unsupervised learning using nonequilibrium thermodynamics". *International conference on machine learning*. pmlr. 2015, 2256–2265 2.

[SZL*22] SONG, YIZHI, ZHANG, ZHIFEI, LIN, ZHE, et al. *ObjectStitch: Generative Object Compositing*. Dec. 2022. DOI: 10.48550/arXiv.2212.00932. arXiv: 2212.00932 [cs]. (Visited on 09/25/2025) 3.

[VHM*21] VERBIN, DOR, HEDMAN, PETER, MILDENHALL, BEN, et al. "Ref-nerf: Structured view-dependent appearance for neural radiance fields". *arXiv preprint arXiv:2112.03907* (2021) 1, 6.

[VMG*25] VIOLANTE, NICOLÁS, MEULEMAN, ANDRÉAS, GAUTHIER, ALBAN, et al. "Splat and Replace: 3D Reconstruction with Repetitive Elements". *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*. SIGGRAPH Conference Papers '25. Association for Computing Machinery, 2025. ISBN: 9798400715402. DOI: 10.1145/3721238.3730727. URL: https://doi.org/10.1145/3721238.3730727 2.

[vPPL*22] Von PLATEN, PATRICK, PATIL, SURAJ, LOZHKOV, ANTON, et al. *Diffusers: State-of-the-art diffusion models*. https://github.com/huggingface/diffusers. 2022 5, 6.

[WCF*24] WINTER, DANIEL, COHEN, MATAN, FRUCHTER, SHLOMI, et al. *ObjectDrop: Bootstrapping Counterfactuals for Photorealistic Object Removal and Insertion*. Mar. 2024. DOI: 10.48550/arXiv.2403.18818. arXiv: 2403.18818 [cs]. (Visited on 09/18/2025) 3.

[WMH*24] WU, RUNDI, MILDENHALL, BEN, HENZLER, PHILIPP, et al. "ReconFusion: 3D Reconstruction with Diffusion Priors". (2024) 3.

[WWA*25] WAN, TEAM, WANG, ANG, AI, BAOLE, et al. "Wan: Open and Advanced Large-Scale Video Generative Models". *arXiv preprint arXiv:2503.20314* (2025) 11.

[WXH*25] WU, SIBO, XU, CONGRONG, HUANG, BINBIN, et al. *GenFusion: Closing the Loop between Reconstruction and Generation via Videos*. Mar. 2025. DOI: 10.48550/arXiv.2503.21219. arXiv: 2503.21219 [cs]. (Visited on 04/04/2025) 3.

[WYLL22] WANG, GUANGCONG, YANG, YINUO, LOY, CHEN CHANGE, and LIU, ZIWEI. "Stylelight: Hdr panorama generation for lighting estimation and editing". *European conference on computer vision*. Springer. 2022, 477–492 2.

[WZT*25] WU, JAY ZHANGJIE, ZHANG, YUXUAN, TURKI, HAITHEM, et al. *Difix3D+: Improving 3D Reconstructions with Single-Step Diffusion Models*. Mar. 2025. DOI: 10.48550/arXiv.2503.01774. arXiv: 2503.01774 [cs]. (Visited on 03/04/2025) 3.

[YDYK23] YE, MINGQIAO, DANELLJAN, MARTIN, YU, FISHER, and KE, LEI. *Gaussian Grouping: Segment and Edit Anything in 3D Scenes*. arXiv:2312.00732 [cs]. Dec. 2023. DOI: 10.48550/arXiv.2312.00732. URL: http://arxiv.org/abs/2312.00732 (visited on 02/26/2024) 2.

[YGZ*22] YANG, BINXIN, GU, SHUYANG, ZHANG, BO, et al. *Paint by Example: Exemplar-based Image Editing with Diffusion Models*. Nov. 2022. DOI: 10.48550/arXiv.2211.13227. arXiv: 2211.13227 [cs]. (Visited on 07/11/2025) 3.

[YHXS25] YU, MARK, HU, WENBO, XING, JINBO, and SHAN, YING. *TrajectoryCrafter: Redirecting Camera Trajectory for Monocular Videos via Diffusion Models*. Mar. 2025. DOI: 10.48550/arXiv.2503.05638. arXiv: 2503.05638 [cs]. (Visited on 03/13/2025) 3.

[YHZ24] YE, KEYANG, HOU, QIMING, and ZHOU, KUN. "3D Gaussian Splatting with Deferred Reflection". *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers '24*. July 2024, 1–10. DOI: 10.1145/3641519.3657456. arXiv: 2404.18454 [cs]. (Visited on 06/20/2025) 2, 6–10.

[YYZ*24] YING, HAIYANG, YIN, YIXUAN, ZHANG, JINZHI, et al. "OmniSeg3D: Omniversal 3D Segmentation via Hierarchical Contrastive Learning". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024 2.

[YZL*23] YE, HU, ZHANG, JUN, LIU, SIBO, et al. *IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models*. Aug. 2023. DOI: 10.48550/arXiv.2308.06721. arXiv: 2308.06721 [cs]. (Visited on 07/24/2024) 2.

[ZCJ*24] ZHOU, SHIJIE, CHANG, HAORAN, JIANG, SICHENG, et al. "Feature 3DGS: Supercharging 3D Gaussian Splatting to Enable Distilled Feature Fields". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2024, 21676–21685 2.

[ZDG*24] ZENG, ZHENG, DESCHAINTRE, VALENTIN, GEORGIEV, ILIYAN, et al. "RGB↔X: Image Decomposition and Synthesis Using Material- and Lighting-Aware Diffusion Models". *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers '24*. Denver CO USA: ACM, July 2024, 1–11. ISBN: 979-8-4007-0525-0. DOI: 10.1145/3641519.3657445. (Visited on 07/18/2024) 2, 4.

[ZDL*23] ZHANG, BO, DUAN, YUXUAN, LAN, JUN, et al. "Controlcom: Controllable image composition using diffusion model". *arXiv preprint arXiv:2308.10040* (2023) 3.

[ZFG*25] ZHANG, ZITIAN, FORTIER-CHOUINARD, FRÉDÉRIC, GARON, MATHIEU, et al. *ZeroComp: Zero-shot Object Compositing from Image Intrinsics via Diffusion*. Jan. 2025. DOI: 10.48550/arXiv.2410.08168. arXiv: 2410.08168 [cs]. (Visited on 07/10/2025) 2.

[ZGF*24] ZHENG, PENG, GAO, DEHONG, FAN, DENG-PING, et al. "Bilateral Reference for High-Resolution Dichotomous Image Segmentation". *CAAI Artificial Intelligence Research* 3 (2024), 9150038 4, 5.

[ZIE*18] ZHANG, RICHARD, ISOLA, PHILLIP, EFROS, ALEXEI A, et al. "The unreasonable effectiveness of deep features as a perceptual metric". *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, 586–595 6, 7, 11.

[ZLL*25] ZHAI, HONGJIA, LI, HAI, LI, ZHENZHE, et al. "PanoGS: Gaussian-based Panoptic Segmentation for 3D Open Vocabulary Scene Understanding". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2025, 14114–14124 2.

[ZRA23] ZHANG, LVMIN, RAO, ANYI, and AGRAWALA, MANEESH. "Adding Conditional Control to Text-to-Image Diffusion Models". *IEEE International Conference on Computer Vision (ICCV)*. 2023 2.

[ZYW*24] ZHU, XUENING, YI, RENJIAO, WEN, XIN, et al. *Relighting Scenes with Object Insertions in Neural Radiance Fields*. June 2024. DOI: 10.48550/arXiv.2406.14806. arXiv: 2406.14806 [cs]. (Visited on 10/29/2025) 6.