MV-CoLight: Efficient Object Compositing with Consistent Lighting and Shadow Generation

Kerui Ren[1,2] **Jiayang Bai**[3] **Linning Xu**[4] **Lihan Jiang**[2,5]
Jiangmiao Pang[2] **Mulin Yu**[2*] **Bo Dai**[6*]
[1]Shanghai Jiao Tong University, [2]Shanghai Artificial Intelligence Laboratory,
[3]Nanjing University, [4]The Chinese University of Hong Kong,
[5]University of Science and Technology of China, [6]The University of Hong Kong

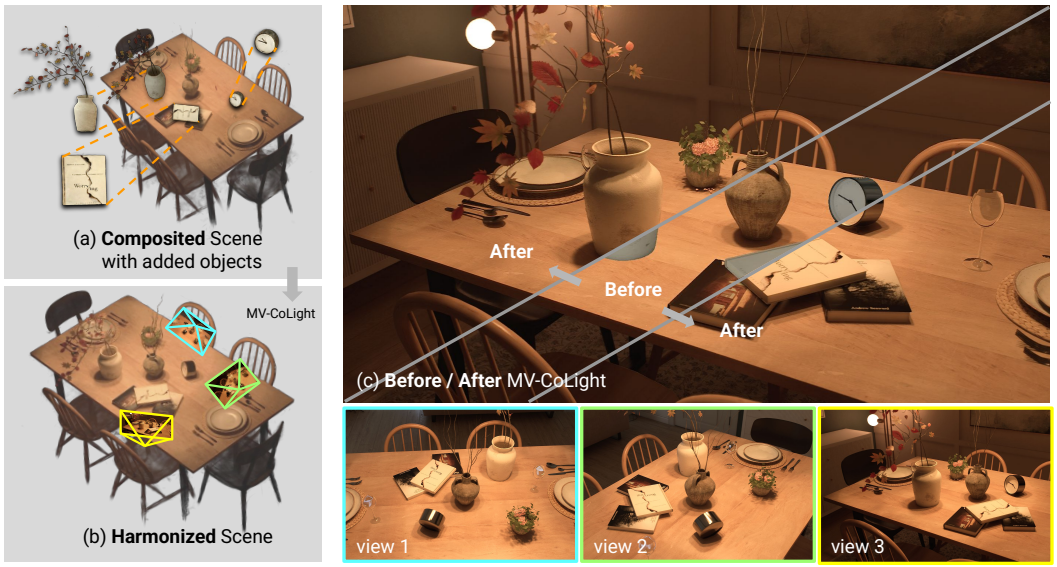

Figure 1: Illustration of our object compositing pipeline with harmonization and relighting using MV-CoLight. In (a), we show a composite scene with visually inconsistent inserted objects. Applying our MV-CoLight method in (b), we generate realistic lighting, shadows, and harmonious integration of objects into the 3D scene. Panel (c) highlights clear visual differences before and after harmonization, accompanied by consistent novel view renderings below. Explore more demos on our project page: **https://city-super.github.io/mvcolight/**.

Abstract

Object compositing offers significant promise for augmented reality (AR) and embodied intelligence applications. Existing approaches predominantly focus on single-image scenarios or intrinsic decomposition techniques, facing challenges with multi-view consistency, complex scenes, and diverse lighting conditions. Recent inverse rendering advancements, such as 3D Gaussian and diffusion-based methods, have enhanced consistency but are limited by scalability, heavy data requirements, or prolonged reconstruction time per scene. To broaden its applicability, we introduce MV-CoLight, a two-stage framework for illumination-consistent object compositing in both 2D images and 3D scenes. Our novel feed-forward architecture models lighting and shadows directly, avoiding the iterative biases of diffusion-based methods. We employ a Hilbert curve–based mapping to align

*Corresponding author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

2D image inputs with 3D Gaussian scene representations seamlessly. To facilitate training and evaluation, we further introduce a large-scale 3D compositing dataset. Experiments demonstrate state-of-the-art harmonized results across standard benchmarks and our dataset, as well as casually captured real-world scenes demonstrate the framework's robustness and wide generalization.

1 Introduction

Object compositing in 3D scenes remains a formidable challenge due to the interplay of color harmonization, shadow synthesis, light transport simulation, and multi-view consistency, all of which must be addressed to achieve photorealistic integration. This capability is fundamental to AR, robotics, and interactive media, where realism directly impacts user immersion and perception.

Early object compositing research focuses primarily on isolated subtasks like scene relighting [54, 17], shadow generation [24, 25], and color harmonization [7, 12], yielding promising yet fragmented solutions. However, The transition toward unified frameworks reveals intricate couplings between these components, necessitating adherence to physical principles governing light transport and occlusion phenomena. Diffusion-based pipelines such as ObjectStitch [38] and ControlCom [53] attempt single-image object insertion by synthesizing harmonious lighting and shadows within a background bounding box, but their reliance on stochastic sampling and the lack of large-scale, high-quality compositing datasets limit their robustness and generalization in real-world scenarios.

In this work, we tackle the problem of seamlessly inserting novel objects into static 3D scenes captured from multiple viewpoints. Our goal is to relight each object so that its appearance, including ambient illumination, surface reflections, and cast shadows, matches the lighting of the scene, while also modeling the reciprocal effects of the object on its surroundings (e.g. secondary shadows and interreflections). We introduce MV-CoLight, a unified framework that preserves both geometric fidelity and photorealism across views by learning and enforcing lighting-consistent priors at both the image and scene levels. MV-CoLight adopts a two-stage training pipeline. In the 2D object compositing stage, we train a feed-forward model to capture scene-specific lighting characteristics, including background shadows and indirect illumination, from individual images. In the 3D object compositing stage, we transform these learned features into a 3D Gaussian representation using 3D Gaussian splatting [19], ordering them via a Hilbert curve to ensure spatial coherence and enforce multi-view consistency. Leveraging recent advances in video-level instance segmentation and 3D-aware object insertion, our framework effectively eliminates common 2D mask artifacts while achieving efficient inference (0.07s per frame) without compromising stability or visual quality.

To support training and evaluation, we introduce a large-scale synthetic dataset of over 480k composite scenes rendered in Blender. Each scene features a table from the Digital Twin Catalog [8], augmented with Poly Haven HDR environment maps and materials [30], and additional light sources for varied illumination. We render 16 uniformly sampled RGB views per scene, along with depth maps and segmentation masks. To simulate realistic compositing challenges, we mix foreground and background layers under different lighting conditions, creating deliberate lighting inconsistencies for training and evaluation. Further implementation details are provided in the supplementary material.

Our main contributions are as follows: 1) a feed-forward architecture for multi-view object compositing that, unlike diffusion-based alternatives, offers improved computational efficiency and robustness with high visual quality; 2) a two-stage training framework that connects 2D object compositing with 3D Gaussian color fields via a Hilbert curve ordering mechanism, thereby enforcing geometrically consistent illumination priors and coherent multi-view shadows; and 3) curate a large-scale benchmark of over 480 K annotated multi-view scenes under varying lighting conditions, and demonstrate that our method achieves state-of-the-art performance across several public datasets.

2 Related Works

Object compositing, the seamless integration of foreground objects into background scenes, is a fundamental task in both image editing and 3D graphics. In the following, we briefly discuss three principal paradigms that have guided existing solutions.

Multi-Task Decomposition Approach. Object compositing generally involves addressing three challenges, including color harmonization, relighting, and shadow generation. Below, we briefly review related works in these areas. Color harmonization has evolved from classical low-level techniques using color statistics and gradient adjustments [20, 32, 39, 44] to learning-based methods [37, 13, 28, 12, 14, 49, 6] powered by large-scale datasets like iHarmony [7]. Relighting modifies an object's shading while preserving its geometry and material properties. Recent learning-based relighting techniques focused on specific image types, including outdoor scenes [10, 51], portraits [54, 31], and human subjects [16, 50], achieving high-quality results. Shadow generation employs diverse strategies, from using pixel height information to generate diverse lighting effects [36, 35] to GAN-based [42, 56, 45] and generative models [25] that bypass ray-tracing requirements. While recent progress in these subdomains demonstrates improved fidelity, multi-view harmonization and physically grounded shadow synthesis remain open challenges, highlighting the need for holistic frameworks that ensure cross-task and cross-view coherence.

End-to-End Unified Frameworks. Unified end-to-end frameworks for image compositing have emerged in recent studies [38, 3, 53, 40]. ObjectStitch [38] introduces a diffusion-based architecture that concurrently addresses geometry correction, harmonization, shadow generation, and view synthesis. ControlCom [53] further enhances composite fidelity by incorporating a dedicated foreground refinement module. However, these approaches predominantly process single-view inputs. Building on ObjectStitch, MureObjectStitch [3] adopts a multi-reference strategy for multi-perspective compositing, yet it still struggles with inconsistent harmonization when applied to multi-view images from the same scene. In contrast, our work leverages 3D modeling to ensure visual consistency across views, directly addressing these limitations. By integrating 3D priors, our approach simplifies the task to color-mapping transformations for inserted objects. This formulation inherently obviates the need for diffusion-based generative capabilities while necessitating precise per-pixel color transformations. Consequently, we employ a feed-forward network rather than diffusion-based models, which prioritize pixel-level generation and often yield unstable color outputs.

Inverse Rendering Paradigm. This approach for object compositing first estimates intrinsic scene properties, such as geometry, materials, and lighting, from input images through inverse rendering [1]. Subsequently, traditional rendering pipelines or neural rendering pipelines are employed to render novel views of the scene with inserted objects. Recent advancements [18, 9, 23] have incorporated 3D scene representations like NeRF [29] and 3D Gaussian Splats [19] within neural rendering pipelines. The emergence of large-scale image generative models [22, 52, 21] has recently revitalized inverse rendering research. The RGB$\leftrightarrow$X [52] framework first trains an image diffusion model to estimate G-buffers from object and scene data. It then composites synthetic objects into these estimated channels and employs a diffusion model to generate final images with consistent lighting and shadow effects. However, such methods demand extensive high-quality datasets with fully paired intrinsic properties to achieve robust generalization capabilities, which poses significant challenges for real-world environment applications.

3 Methods

In this work, we focus on efficiently synthesizing consistent lighting and shadows to harmonize scenes. Formally, given a background scene and a foreground object, we assume that the inserted object and the background scene are entirely aligned and our task is to produce multi-view renderings that insert, harmonize, and relight the object under novel illumination while maintaining overall coherence, which presents an essential requirement for AR and embodied-intelligence applications that demand real-time, view-consistent integration.

Fig. 2 illustrates our two-stage framework. (1) We begin with an inharmonious scene and convert it into a pixel-aligned 3D Gaussian representation (Sec.3.1). (2) Each image is then processed independently by a transformer-based network for single-view object compositing (Sec.3.2). (3) To achieve multi-view consistency, we concatenate the extracted pixel-wise features with Gaussian-wise features, order them along a Hilbert curve [15], and decode them with a second transformer to predict harmonious Gaussian color attributes (Sec.3.3). (4) Finally, Sec.3.4 describes the loss functions that drive our two-stage training. A brief introduction to Gaussian splatting and the Hilbert curve is provided in the supplementary materials.

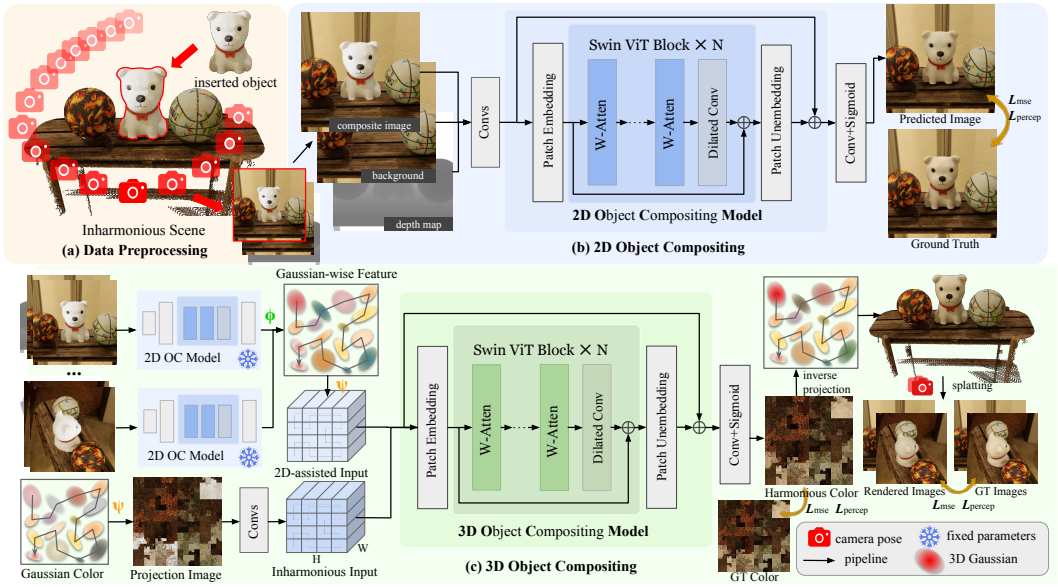

Figure 2: Pipeline of MV-CoLight. In (a), we insert a white puppy as the composite object onto the table between basketballs, and render multi-view inharmonious images, background-only images, and depth maps using a camera trajectory moving from distant to close-up positions. Subsequently in (b), we input a single-view data into the 2D object compositing model, which processes the data through multiple Swin Transformer blocks to output the harmonized result. Finally in (c), we project the multi-view features from 2D models into Gaussian space via $\Phi(\cdot)$, combine them with the original inharmonious Gaussian colors projected into 2D Gaussian color space through $\Psi(\cdot)$, and then feed them into the 3D object compositing model. The model outputs harmonized Gaussian colors and computes rendering loss by incorporating Gaussian shape attributes.

3.1 Data Preprocessing

Begin with a composed 3D scene, obtained via synthesis, 3D scanning, or multiview reconstruction pipeline (e.g., [34, 43]), we place a set of cameras orbiting the scene center to obtain multi-view composite images, background-only images, and depth maps. From each view's images, camera poses, and depth data, we build point maps including 3D positions and colors, then randomly sample a fixed number M of points to initialize the 3D Gaussian model $\mathcal{G}'$. During optimization, we fix each Gaussian's opacity at 1 and adjust only its shape parameters. As illustrated in Fig. 3, we organize 3D Gaussian primitives, each tied to a unique training pixel, into spatially coherent patches by mapping their centers along a space-filling Hilbert curve[15], denoted as mapping $\Phi(\cdot)$.

3.2 2D Object Compositing

Given an inharmonious composite image $I \in \mathbb{R}^{3\times H\times W}$, its background reference $G \in \mathbb{R}^{3\times H\times W}$, and depth map $D \in \mathbb{R}^{1\times H\times W}$, we form the input tensor $\{I, G, D\}$ and feed it into our 2D object compositing network:

$$\hat{H} = \mathcal{M}_{2d}(\{I, G, D\}; \theta_{2d}) \tag{1}$$

where $\hat{H}$ is the predicted harmonized image and θ_{2d} are the network parameters. Following Grounded DINO [26] and DINO-X [33], we adopt the Swin Transformer as our backbone. Its shifted-window attention mechanism achieves linear computational complexity with respect to image size, while the hierarchical pyramid structure ensures high hardware efficiency. Compared to the original transformer architecture, swin transformer delivers superior performance across a variety of vision tasks, yet requires fewer FLOPS and parameters under the same input resolution.

Specifically, $\mathcal{M}_{2d}$ begins with several 3×3 convolutions to extract shallow features, which are then partitioned into non-overlapping $P \times P$ patches and fed into L Swin Transformer layers [27]. For each layer $i \in \{1, \ldots, L\}$, the input F_{i-1} is normalized, processed by window-based multi-head self-attention with a residual connection, renormalized, passed through an MLP with a second residual

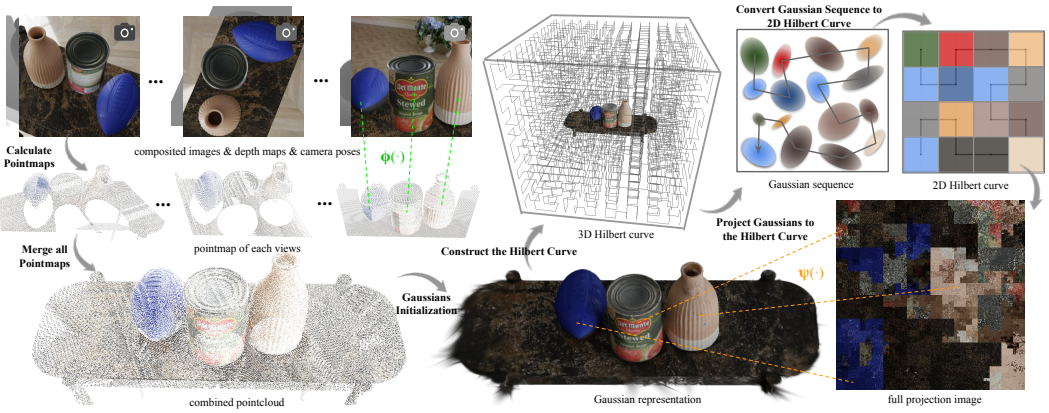

Figure 3: Mapping multi-view observations into a 2D Hilbert-ordered Gaussian color map. Starting from inharmonious multi-view images, depth maps, and camera poses, we compute per-view point maps and randomly sample M points to initialize 3D Gaussian primitives, which we then optimize to fit the scene. Next, we construct a 3D Hilbert curve through the Gaussian centers and assign each primitive to its nearest curve point, yielding an ordered 1D sequence. Finally, we fold this sequence into a 2D grid along a 2D Hilbert curve, producing a spatially coherent projection in which each pixel encodes the color of its corresponding Gaussian.

skip, and output as F_i, as summarized by

$$\begin{aligned} \hat{F}_i &= \text{W-Atten}\big(\text{LN}(F_{i-1})\big) + F_{i-1}, \\ F_i &= \text{MLP}\big(\text{LN}(\hat{F}_i)\big) + \hat{F}_i. \end{aligned} \tag{2}$$

At inference time, $\mathcal{M}_{2d}$ is applied independently to each input. We extract the output feature maps $F \in \mathbb{R}^{m\times n\times H\times W}$ from the final attention block, where m is the number of views and n is the feature dimension. These features are then transformed into 3D Gaussians via the mapping function $\Phi(\cdot)$ for downstream 3D compositing network (Sec. 3.3). See supplementary material for more details.

3.3 3D Object Compositing

For 3D object compositing, we seek an illumination-consistent Gaussian model $\mathcal{G}$ that preserves each primitive's 3D position (x, y, z), scale, and rotation from the inharmonious model $\mathcal{G}'$, but updates only its color attributes $\mathcal{C}'$. Thus, we freeze all positional and geometric parameters and learn a color-only mapping. Inspired by recent advancements [47, 5] in processing point clouds with transformers, like Point Transformer V3, we use a mapping Ψ that first linearizes the sparse 3D Gaussians into a 1D sequence via a 3D Hilbert curve and then arranges them into a 2D grid by the inverse 2D Hilbert curve. We concatenate these Gaussian-wise features, combining the transformed 2D colors and the $\mathcal{M}_{2d}$ features via $\Phi(\cdot)$, and feed them into our 3D compositing network:

$$I_{\hat{\mathcal{C}}} \;=\; \mathcal{M}_{3d}\big(\{\Psi(\Phi(F)),\, \Psi(\mathcal{C}')\};\, \theta_{3d}\big),$$

where θ_{3d} are the network parameters and $\mathcal{C}'$ are the original inharmonious Gaussian colors. The output $I_{\hat{\mathcal{C}}}$ gives harmonious, view-consistent colors, which we back-project onto the 3D Gaussians to complete the compositing. Notably, $\mathcal{M}_{3d}$ adopts the similar architectural designs as $\mathcal{M}_{2d}$, differing only in its input and output dimensions. Please refer to supplementary material for more details.

3.4 Loss Design

During the two-stage training process, we employ similar loss function design, utilizing mean-square error loss $\mathcal{L}_{mse}$ loss and perceptual loss $\mathcal{L}_p$ to optimize the object compositing models:

$$\mathcal{L}_{2d} = \mathcal{L}_{mse}(\hat{H}, H) + \lambda\mathcal{L}_p(\hat{H}, H) \tag{3}$$

$$\mathcal{L}_{3d} = \beta(\mathcal{L}_{mse}(I_{\hat{\mathcal{C}}}, \Psi(\mathcal{C})) + \lambda\mathcal{L}_p(I_{\hat{\mathcal{C}}}, \Psi(\mathcal{C}))) + \frac{(1-\beta)}{n}\sum_{i=1}^{n}(\mathcal{L}_{mse}(\hat{H}_i, H_i) + \lambda\mathcal{L}_p(\hat{H}_i, H_i)) \tag{4}$$

Table 1: Single-view quantitative performance on our purposed dataset and the Objects With Lighting dataset [41]. We report visual quality metrics, inference time and memory storage, highlighting the **best** and second-best in each category. Our* and Our† denote our method without depth input and without both depth and background input, respectively.

Dataset		Simple Synthetic Scene			Complex Synthetic Scene			Objects With Lighting			Time↓	Memory↓
Paradigm	Method	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓		
Diffusion-based	LumiNet [48]	16.94	0.614	0.287	19.94	0.671	0.274	17.15	0.781	0.222	23.82s	13.79G
Feed-forward	GPT-4o [4]	14.60	0.418	0.437	15.13	0.369	0.415	12.14	0.479	0.351	1.36m	-
Feed-forward	Ours†	28.35	0.957	0.031	29.61	0.947	0.029	27.48	0.945	0.051	0.07s	32.89M
Feed-forward	PCT-Net [12]	22.58	0.912	0.055	25.26	0.931	0.035	25.08	0.921	0.066	**0.03s**	**18.4M**
Diffusion-based	Objectstitch [38]	19.14	0.770	0.193	21.82	0.788	0.170	21.15	0.831	0.176	4.54s	5.24G
Diffusion-based	ControlCom [53]	18.85	0.765	0.209	19.88	0.771	0.185	19.75	0.811	0.189	4.63s	10.94G
Diffusion-based	RGB↔X [52]	12.28	0.428	0.368	12.91	0.507	0.296	11.28	0.503	0.422	19.71s	10.68G
Diffusion-based	IC-Light [54]	17.66	0.659	0.217	20.87	0.679	0.190	18.22	0.774	0.200	1.25s	1.60G
Feed-forward	Ours*	29.11	0.959	0.030	30.00	0.951	0.027	28.18	0.945	0.050	0.07s	32.92M
Feed-forward	Ours	**29.65**	**0.961**	**0.029**	**30.20**	**0.953**	**0.027**	**28.75**	**0.946**	**0.049**	0.07s	32.94M

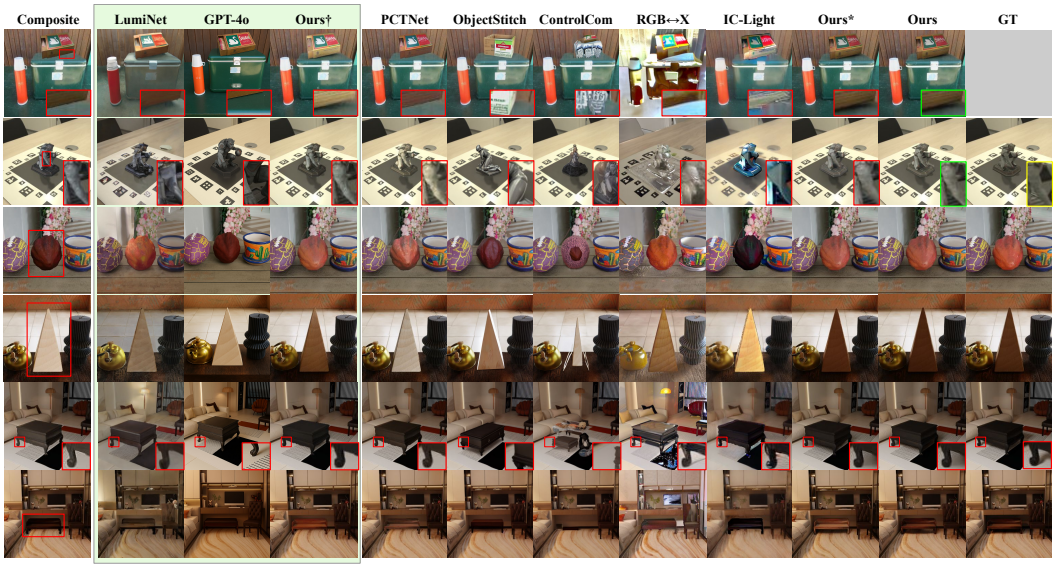

Figure 4: Single-view qualitative comparison with SOTA methods [48, 4, 12, 38, 53, 52, 54] on our proposed dataset and public datasets [53, 41], with differences highlighted via colored patches. Compared to existing baselines, our method successfully generates illumination consistent with the background and physically plausible shadows while decoupling highlights from inserted objects, demonstrating generalization capabilities on out-of-domain datasets. The method in the green box does not incorporate background images as input, whereas the others do.

where H_i and $\hat{H}_i$ denote the ground truth images and the rendered images from the harmonized Gaussian $\mathcal{G}'$, which is composed of $\mathcal{C}'$ and shape parameter from $\mathcal{G}$, λ and β are the hyper-parameter and set as 0.05 and 0.5 by default.

4 Experiments

4.1 Experimental Setup

Datasets and Metrics. Our proposed dataset contains simple synthetic, complex synthetic, and real captured scenes. From the simple synthetic set, 50 scenes are randomly held out for evaluation, while the rest are used for training. In contrast, both the complex synthetic and real captured scenes are solely used for evaluation. Besides, we evaluate our method on two public benchmarks, FOSCom [53] and Objects With Lighting (OWL) [41]. For 2D object compositing, we test on 640 scenes from FOSCom, 72 scenes from OWL, and 58 challenging scenes from our proposed dataset. For 3D object compositing, we report results on 50 simple and 8 complex synthetic scenes from our dataset,

Table 2: Multi-view quantitative performance on our purposed dataset and real captured scenes. We report visual quality metrics, inference time (Gaussian training time *#Train*), highlighting the **best** and second-best in each category. Our* and Our† denote our method without depth input and without both depth and background input, respectively.

Dataset		Simple Synthetic Scene			Complex Synthetic Scene			Real Captured Scene			Time↓ (#Train)
Paradigm	Method	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	
Diffusion-based	LumiNet [48]	17.15	0.573	0.304	18.45	0.663	0.222	20.05	0.770	0.198	6.31m (-)
Feed-forward	GPT-4o [4]	14.57	0.375	0.445	15.11	0.366	0.411	14.34	.473	0.406	21.40m (-)
Feed-forward	Ours†	28.96	0.955	0.033	29.32	0.946	0.029	25.88	0.925	0.041	1.07s (1.08m)
Feed-forward	PCT-Net [12]	22.97	0.908	0.057	25.19	0.927	0.035	23.39	0.824	0.103	**0.47s** (-)
Diffusion-based	Objectstitch [38]	19.12	0.726	0.217	21.84	0.792	0.163	18.43	0.785	0.193	1.21m (-)
Diffusion-based	ControlCom [53]	18.95	0.722	0.231	19.60	0.773	0.181	18.54	0.778	0.199	1.23m (-)
Diffusion-based	RGBX [52]	12.71	0.417	0.360	12.69	0.504	0.304	13.68	0.594	0.312	5.26m (-)
Diffusion-based	IC-Light [54]	17.94	0.596	0.242	20.60	0.689	0.183	20.23	0.718	0.233	19.36s (-)
Inverse Rendering	GS-IR [23]	15.56	0.742	0.134	16.81	0.664	0.249	15.92	0.699	0.265	17.62s (57.14m)
Inverse Rendering	GI-GS [2]	18.97	0.808	0.126	16.56	0.674	0.310	16.07	0.716	0.234	16.69s (1.43h)
Inverse Rendering	IRGS [11]	17.79	0.688	0.237	21.04	0.702	0.291	20.19	0.744	0.215	9.72m (3.02h)
Feed-forward	Ours*	29.73	0.958	0.031	29.51	0.949	0.028	26.10	0.926	0.041	1.07s (1.08m)
Feed-forward	Ours	**30.29**	**0.960**	**0.030**	**30.13**	**0.952**	**0.027**	**26.39**	**0.927**	**0.040**	1.08s (1.08m)

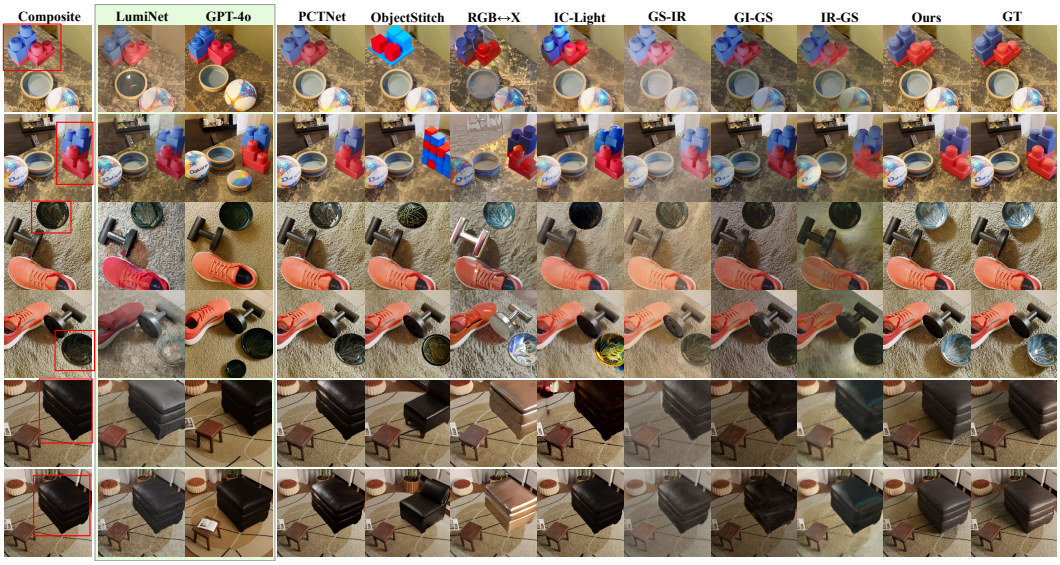

Figure 5: Multi-view qualitative comparison with SOTA methods [48, 4, 12, 38, 53, 52, 54, 23, 2, 11] on our proposed dataset and real captured scenes, with differences highlighted via colored patches. Our method synthesizes plausible illumination and shadows while ensuring multi-view consistency. The method in the green box does not incorporate background images as input, whereas the others do.

with another two real captured scenes. All images are center-cropped and rescaled to 256×256 for uniform comparison. Performance is quantified using PSNR, SSIM [46], and LPIPS [55]. Since each ground-truth image embodies only one physically plausible lighting/albedo configuration, perceptual metrics (SSIM and LPIPS) offer additional assessments of structural and visual fidelity than PSNR alone.

Baselines. For 2D object compositing evaluation, we conduct comprehensive comparisons with representative methods: PCTNet [12], ObjectStitch [38], ControlCom [53], RGB↔X [52], ICLight [54], LumiNet [48] and GPT-4o [4]. For 3D object compositing evaluation, we additionally incorporate Gaussian-based inverse rendering method such as GS-IR [23], GI-GS [2], and IRGS [11], establishing a unified benchmark comparing conventional 2D pipelines with emerging 3D-aware methods built upon differentiable rendering frameworks.

Implementation Details. Our model architecture employs a unified Swin Transformer backbone with consistent configurations for both 2D and 3D object compositing tasks. The network processes

Figure 6: We evaluate our method on real-world scenes captured under varying illumination with six cameras arranged in a circular array. On the left, we insert a cake and a black box; on the right, we insert a toy, a mouse, and a backpack. MV-CoLight consistently harmonizes object colors, produces physically plausible lighting interactions, and accurately casts shadows across all viewpoints.

256×256 resolution inputs with an embedding dimension of 96, structured with 3 cascaded transformer blocks. Each block contains 6 successive Swin Transformer layers with 6 parallel attention heads.

We train the 2D object compositing model for 1M iterations with batch size 128 using AdamW (base lr=2e-3, weight decay=0.05, momentum parameters β_1=0.9, β_2=0.95), 10k iteration linear warmup followed by cosine decay to 1e-6, FP16 mixed precision, gradient clipping at 10.0, and an EMA of 0.99. For 3D object compositing model, we reduce the learning rate to 1e-3 and batch size to 32, training for 100k iterations. For trainig time, We train the 2D model from scratch for 15 days, and the 3D model for 3 days with 16 NVIDIA A100 (80 GB) GPUs.

4.2 Performance Analysis

Below we show our method delivers physically plausible lighting and shadows for inserted objects, out-performing both 2D harmonization [48, 4, 12, 38, 53, 52, 54] and Gaussian-based inverse rendering [23, 2, 11] baselines. The approach also generalizes from synthetic training to challenging real-world captures, maintaining photorealism under diverse lighting and materials.

Single-view Harmonized Result. Current image harmonization methods exhibit notable limitations when compositing new objects into a scene, as illustrated in Fig. 4. For example, PCT-Net[12] enforces only color consistency and omits realistic highlights and cast shadows, while RGB-X [52] material estimation yields inaccurate albedo maps that blur illumination and misalign geometry during neural relighting. Diffusion-based frameworks such as ObjectStitch [38] and ControlCom [53] produce visually compelling composites but often distort object shape and texture in the generative process. ICLight [54]'s illumination estimator lacks robustness in cross-domain scenarios, resulting in pronounced appearance artifacts under complex real-world lighting, while the light transport module of LumiNet [48] generates non-physical highlight patterns and jagged shadow boundaries. Even advanced multimodal systems like GPT-4o [4], which improve local lighting coherence, introduce unintended global modifications that undermine overall scene integrity which is particularly hard to be strictly enforced via prompting.

Multi-view Harmonized Result. Multi-view object compositing compounds the inconsistencies of 2D harmonization methods, resulting in visible color shifts and misaligned shadows across viewpoints. Gaussian-based inverse rendering techniques attempt to remedy this by enriching each primitive with material attributes, such as estimated albedo and normals, and estimating an environment map from the background Gaussians to relight the composite. However, their reliance on imperfect decoupling causes specular highlights and shadowed regions from the original images to be treated as textures. As shown in Fig. 5, the result is a conflated relighting effect that blurs the distinction between intrinsic material properties and new environmental illumination, failing to achieve true multi-view coherence. Unlike environment mapping-based methods, our approach directly learns illumination and shadow priors from the multi-view composite scene and transfers them to the inserted object, guaranteeing seamless, view-consistent lighting. By encoding these learned visual cues into Gaussian feature representations and propagating them through our transformer-based 2D-3D pipeline, we maintain spatial coherence and realistic shadowing without explicit environment map estimation. Extensive evaluations on public benchmarks [53, 41] and our own dataset demonstrate that our method outperforms both 2D harmonization and Gaussian-based inverse-rendering baselines in quantitative metrics and visual quality.

Real Scene Harmonized Result. We further assess our method on diverse real-world multi-view captures that diverge markedly from our synthetic training data in terms of lighting complexity and

Figure 7: Visual results of inserting luminous objects. Our method successfully simulates the illumination effects of luminous spheres within the scene environment.

Table 3: Ablation study conducted on the simplified synthetic scenes within our proposed dataset. We report visual quality metrics, inference time and model parameters. For variable control, the corresponding 2D models for 3D object compositing are kept as baseline versions, and their parameters are not counted.

Model	2D Object Compositing Model					3D Object Compositing Model				
Method	PSNR↑	SSIM↑	LPIPS↓	Time↓	Params.↓	PSNR↑	SSIM↑	LPIPS↓	Time↓	Params.↓
baseline	29.65	0.961	0.029	0.07s	2.957M	30.29	0.960	0.030	1.08s	2.957M
transformer block (2)	28.34	0.955	0.032	0.06s	2.085M	28.70	0.953	0.035	0.97s	2.085M
transformer layer (4)	28.39	0.956	0.032	0.06s	2.119M	28.83	0.957	0.034	0.98s	2.118M
embedding dim (60)	27.68	0.951	0.035	0.05s	1.779M	28.11	0.949	0.039	0.81s	1.778M
w/ Linear transform	-	-	-	-	-	28.99	0.951	0.036	1.08s	2.957M
w/ PTv3 feat. extractor	-	-	-	-	-	30.08	0.959	0.031	1.16s	49.16M
w/o 2D OC model	-	-	-	-	-	25.83	0.913	0.051	0.08s	2.957M
w/o depth input	29.11	0.959	0.030	0.07s	2.957M	29.73	0.958	0.031	1.07s	2.957M
w/o background input	28.81	0.958	0.030	0.07s	2.955M	29.44	0.956	0.032	1.07s	2.957M

material detail, as shown in Fig. 6. These scenes feature unpredictable illumination conditions and intricate textures that typically confound traditional harmonization and inverse-rendering pipelines. Nevertheless, our approach consistently produces photorealistic composites, accurately estimating lighting and casting coherent shadows across all viewpoints. This robust performance under uncontrolled, real-world conditions highlights the generalization robustness of our learning-based illumination priors.

4.3 Efficiency and Extensibility

Inference Time Comparison. For 2D object compositing, our feedforward architecture enables inference times as short as about 0.07 seconds, significantly outperforming diffusion-based methods like LumiNet [48], which require more than 20 seconds to complete the multi-step denoising process (default 50 steps). While GPT-4o [4] remains closed-source, we utilize its web interface to generate results, incurring a latency of several minutes. For 3D object compositing, Gaussian-based inverse rendering methods necessitate both environmental map extraction and material attribute optimization atop pre-trained Gaussians. In contrast, our method achieves harmonized Gaussian color attributes with about 1 second of inference time after a few minutes of scene-specific Gaussian representation learning, demonstrating superior efficiency. Notably, our framework eliminates the need for Gaussian retraining when repositioning inserted objects. This efficiency makes our framework especially well suited for real-time AR and embodied-intelligence applications.

Extentions on Various Illumination Priors. By design, our unified compositing pipeline can also accommodate other illumination priors - whether HDR environment maps, learned light distributions, or discrete emitters, with similar training and inference pipeline. This challenging capability has been largely neglected by prior works. As a demonstrative extension in Fig. 7, we focus here on inserting new light sources to dynamically relight the scene. To support emissive-object compositing with true multi-view consistency, our method estimates light propagation through the existing 3D Gaussian geometry, capturing how point or area emitters illuminate surrounding surfaces, and computes secondary shadows and interreflections to generate physically plausible shading on all objects. This extension underscores the flexibility of our approach and its applicability to a wide range of illumination scenarios.

4.4 Ablation Study

We perform a series of ablations to isolate the factors driving our 2D and 3D compositing pipelines, the results are shown in Tab. 3. First, we vary the number of swin transformer blocks, layers and the feature-embedding dimension. We find that embedding size has a greater impact on final performance than the depth of the transformer stack, with a modest drop in harmonization quality when reducing the block number or layer number. Next, we examine the 2D compositing inputs, removing the background image prevented reliable object placement, while omitting the depth map eliminated essential geometric priors. We find that both scenarios degrade the model's ability to learn illumination conditions. When we feed these 2D features into the 3D network, excluding them entirely still allowed plausible color matching but produced unrealistic shadows and highlights, underscoring the importance of 2D illumination cues. Introducing the Hilbert curve reordering further accelerate training convergence and improve visual quality by preserving Gaussian color locality in 2D Hilbert space. Furthermore, a comparison with the use of PTv3 [47] to extract Gaussian features also indicates that our method based on the Hilbert curve is more lightweight and effective.

5 Limitation

While our approach achieves superior performance across diverse synthetic and real-world scenes, several limitations persist: (a) Color bias in real-world scenes. Trained predominantly on large-scale synthetic data, the model occasionally exhibits color discrepancies when applied to certain real-world environments. (b) Physically inconsistent illumination. Due to the absence of strict physical constraints, deviations in specular highlight positions and shadow directions may arise under complex lighting conditions. (c) Gaussian representation limitations. Errors in Gaussian parameterization, constrained by their inherent capacity to model complex scene details, can propagate to degrade harmonization quality. To address these problems, future work may focus on: (a) Integration of physical constraints. Enhancing shadow consistency by estimating light source positions and constraining shadow regions. (b) 4D scene harmonization. Extending harmonization to dynamic scenes to enable consistent movement of inserted objects within dynamic environments.

6 Conclusion

In this paper, we introduce MV-CoLight, a two-stage framework that seamlessly combines a 2D feed-forward harmonization network with a 3D Gaussian-based compositing model to deliver efficient, view-consistent object insertion. In the first stage, our 2D network rapidly learns per-view color and illumination alignment; in the second, the 3D Gaussian fields enforce geometric and lighting coherence across viewpoints, producing realistic shadows and reflections with minimal runtime overhead. Extensive experiments on both synthetic and real-world benchmarks show that MV-CoLight outperforms state-of-the-art 2D and 3D baselines in visual fidelity and consistency. To drive further progress, we also introduce a new large-scale multi-view compositing dataset with photorealistic accurate annotations. Finally, we demonstrate that our pipeline naturally generalizes to additional lighting effects, underscoring its versatility for broader applications.

7 Acknowledgments

This work was funded in part by the National Key R&D Program of China (2022ZD0160201), Shanghai Artificial Intelligence Laboratory, the National Natural Science Foundation of China (Grant No. 62502247) and the HKU Startup Fund.

References

[1] Harry Barrow, J Tenenbaum, A Hanson, and E Riseman. Recovering intrinsic scene characteristics. *Comput. vis. syst*, 2(3-26):2, 1978.

[2] Hongze Chen, Zehong Lin, and Jun Zhang. Gi-gs: Global illumination decomposition on gaussian splatting for inverse rendering. *arXiv preprint arXiv:2410.02619*, 2024.

[3] Jiaxuan Chen, Bo Zhang, and Li Niu. Mureobjectstitch: Multi-reference image composition. *arXiv preprint arXiv:2411.07462*, 2024.

[4] Sixiang Chen, Jinbin Bai, Zhuoran Zhao, Tian Ye, Qingyu Shi, Donghao Zhou, Wenhao Chai, Xin Lin, Jianzong Wu, Chao Tang, et al. An empirical study of gpt-4o image generation capabilities. *arXiv preprint arXiv:2504.05979*, 2025.

[5] Wanli Chen, Xinge Zhu, Guojin Chen, and Bei Yu. Efficient point cloud analysis using hilbert curve. In *European Conference on Computer Vision*, pages 730–747. Springer, 2022.

[6] Wenyan Cong, Xinhao Tao, Li Niu, Jing Liang, Xuesong Gao, Qihao Sun, and Liqing Zhang. High-resolution image harmonization via collaborative dual transformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18470–18479, 2022.

[7] Wenyan Cong, Jianfu Zhang, Li Niu, Liu Liu, Zhixin Ling, Weiyuan Li, and Liqing Zhang. Dovenet: Deep image harmonization via domain verification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8394–8403, 2020.

[8] Zhao Dong, Ka Chen, Zhaoyang Lv, Hong-Xing Yu, Yunzhi Zhang, Cheng Zhang, Yufeng Zhu, Stephen Tian, Zhengqin Li, Geordie Moffatt, et al. Digital twin catalog: A large-scale photorealistic 3d object digital twin dataset. *arXiv preprint arXiv:2504.08541*, 2025.

[9] Jian Gao, Chun Gu, Youtian Lin, Zhihao Li, Hao Zhu, Xun Cao, Li Zhang, and Yao Yao. Relightable 3d gaussians: Realistic point cloud relighting with brdf decomposition and ray tracing. In *European Conference on Computer Vision*, pages 73–89. Springer, 2024.

[10] David Griffiths, Tobias Ritschel, and Julien Philip. Outcast: Outdoor single-image relighting with cast shadows. In *Computer Graphics Forum*, volume 41, pages 179–193. Wiley Online Library, 2022.

[11] Chun Gu, Xiaofei Wei, Zixuan Zeng, Yuxuan Yao, and Li Zhang. Irgs: Inter-reflective gaussian splatting with 2d gaussian ray tracing. *arXiv preprint arXiv:2412.15867*, 2024.

[12] Julian Jorge Andrade Guerreiro, Mitsuru Nakazawa, and Björn Stenger. Pct-net: Full resolution image harmonization using pixel-wise color transformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5917–5926, 2023.

[13] Zonghui Guo, Dongsheng Guo, Haiyong Zheng, Zhaorui Gu, Bing Zheng, and Junyu Dong. Image harmonization with transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 14870–14879, 2021.

[14] Zonghui Guo, Haiyong Zheng, Yufeng Jiang, Zhaorui Gu, and Bing Zheng. Intrinsic image harmonization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16367–16376, June 2021.

[15] David Hilbert. *Dritter Band: Analysis· Grundlagen der Mathematik· Physik Verschiedenes: Nebst Einer Lebensgeschichte*. Springer-Verlag, 2013.

[16] Chaonan Ji, Tao Yu, Kaiwen Guo, Jingxin Liu, and Yebin Liu. Geometry-aware single-image full-body human relighting. In *European Conference on Computer Vision*, pages 388–405. Springer, 2022.

[17] Haian Jin, Yuan Li, Fujun Luan, Yuanbo Xiangli, Sai Bi, Kai Zhang, Zexiang Xu, Jin Sun, and Noah Snavely. Neural gaffer: Relighting any object via diffusion. *Advances in Neural Information Processing Systems*, 37:141129–141152, 2025.

[18] Haian Jin, Isabella Liu, Peijia Xu, Xiaoshuai Zhang, Songfang Han, Sai Bi, Xiaowei Zhou, Zexiang Xu, and Hao Su. Tensoir: Tensorial inverse rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 165–174, 2023.

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

[20] Jean-Francois Lalonde and Alexei A Efros. Using color compatibility for assessing image realism. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE, 2007.

[21] Ruofan Liang, Zan Gojcic, Huan Ling, Jacob Munkberg, Jon Hasselgren, Zhi-Hao Lin, Jun Gao, Alexander Keller, Nandita Vijaykumar, Sanja Fidler, et al. Diffusionrenderer: Neural inverse and forward rendering with video diffusion models. *arXiv preprint arXiv:2501.18590*, 2025.

[22] Ruofan Liang, Zan Gojcic, Merlin Nimier-David, David Acuna, Nandita Vijaykumar, Sanja Fidler, and Zian Wang. Photorealistic object insertion with diffusion-guided inverse rendering. In *European Conference on Computer Vision*, pages 446–465. Springer, 2024.

[23] Zhihao Liang, Qi Zhang, Ying Feng, Ying Shan, and Kui Jia. Gs-ir: 3d gaussian splatting for inverse rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21644–21653, 2024.

[24] Daquan Liu, Chengjiang Long, Hongpan Zhang, Hanning Yu, Xinzhi Dong, and Chunxia Xiao. Arshadowgan: Shadow generative adversarial network for augmented reality in single light scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8139–8148, 2020.

[25] Qingyang Liu, Junqi You, Jianting Wang, Xinhao Tao, Bo Zhang, and Li Niu. Shadow generation for composite image using diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8121–8130, 2024.

[26] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pages 38–55. Springer, 2024.

[27] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

[28] Quanling Meng, Liu Qinglin, Zonglin Li, Xiangyuan Lan, Shengping Zhang, and Liqiang Nie. High-resolution image harmonization with adaptive-interval color transformation. *Advances in Neural Information Processing Systems*, 37:13769–13793, 2025.

[29] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

[30] Poly Haven. Poly haven. `https://polyhaven.com/`.

[31] Pramod Rao, Gereon Fox, Abhimitra Meka, Mallikarjun BR, Fangneng Zhan, Tim Weyrich, Bernd Bickel, Hanspeter Pfister, Wojciech Matusik, Mohamed Elgharib, et al. Lite2relight: 3d-aware single image portrait relighting. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–12, 2024.

[32] Erik Reinhard, Michael Adhikhmin, Bruce Gooch, and Peter Shirley. Color transfer between images. *IEEE Computer graphics and applications*, 21(5):34–41, 2001.

[33] Tianhe Ren, Yihao Chen, Qing Jiang, Zhaoyang Zeng, Yuda Xiong, Wenlong Liu, Zhengyu Ma, Junyi Shen, Yuan Gao, Xiaoke Jiang, et al. Dino-x: A unified vision model for open-world object detection and understanding. *arXiv preprint arXiv:2411.14347*, 2024.

[34] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[35] Yichen Sheng, Yifan Liu, Jianming Zhang, Wei Yin, A Cengiz Oztireli, He Zhang, Zhe Lin, Eli Shechtman, and Bedrich Benes. Controllable shadow generation using pixel height maps. In *European Conference on Computer Vision*, pages 240–256. Springer, 2022.

[36] Yichen Sheng, Jianming Zhang, and Bedrich Benes. Ssn: Soft shadow network for image compositing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4380–4390, 2021.

[37] Konstantin Sofiiuk, Polina Popenova, and Anton Konushin. Foreground-aware semantic representations for image harmonization. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1620–1629, 2021.

[38] Yizhi Song, Zhifei Zhang, Zhe Lin, Scott Cohen, Brian Price, Jianming Zhang, Soo Ye Kim, and Daniel Aliaga. Objectstitch: Object compositing with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18310–18319, 2023.

[39] Kalyan Sunkavalli, Micah K Johnson, Wojciech Matusik, and Hanspeter Pfister. Multi-scale image harmonization. *ACM Transactions on Graphics (TOG)*, 29(4):1–10, 2010.

[40] Gemma Canet Tarrés, Zhe Lin, Zhifei Zhang, Jianming Zhang, Yizhi Song, Dan Ruta, Andrew Gilbert, John Collomosse, and Soo Ye Kim. Thinking outside the bbox: Unconstrained generative object compositing. *arXiv preprint arXiv:2409.04559*, 2024.

[41] Benjamin Ummenhofer, Sanskar Agrawal, Rene Sepulveda, Yixing Lao, Kai Zhang, Tianhang Cheng, Stephan Richter, Shenlong Wang, and German Ros. Objects with lighting: A real-world dataset for evaluating reconstruction and rendering for object relighting. In *2024 International Conference on 3D Vision (3DV)*, pages 137–147. IEEE, 2024.

[42] Lucas Valença, Jinsong Zhang, Michaël Gharbi, Yannick Hold-Geoffroy, and Jean-François Lalonde. Shadow harmonization for realistic compositing. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–12, 2023.

[43] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

[44] Ke Wang, Michaël Gharbi, He Zhang, Zhihao Xia, and Eli Shechtman. Semi-supervised parametric real-world image harmonization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5927–5936, 2023.

[45] Tianyu Wang, Jianming Zhang, Haitian Zheng, Zhihong Ding, Scott Cohen, Zhe Lin, Wei Xiong, Chi-Wing Fu, Luis Figueroa, and Soo Ye Kim. Metashadow: Object-centered shadow detection, removal, and synthesis, 2024.

[46] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[47] Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point transformer v3: Simpler faster stronger. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4840–4851, 2024.

[48] Xiaoyan Xing, Konrad Groh, Sezer Karaoglu, Theo Gevers, and Anand Bhattad. Luminet: Latent intrinsics meets diffusion models for indoor scene relighting. *arXiv preprint arXiv:2412.00177*, 2024.

[49] Ben Xue, Shenghui Ran, Quan Chen, Rongfei Jia, Binqiang Zhao, and Xing Tang. Dccf: Deep comprehensible color filter learning framework for high-resolution image harmonization. In *European conference on computer vision*, pages 300–316. Springer, 2022.

[50] Kanamori Yoshihiro. Relighting humans: occlusion-aware inverse rendering for full-body human images. *ACM Trans. Graph.*, 37:270–1, 2018.

[51] Ye Yu, Abhimitra Meka, Mohamed Elgharib, Hans-Peter Seidel, Christian Theobalt, and William AP Smith. Self-supervised outdoor scene relighting. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*, pages 84–101. Springer, 2020.

[52] Zheng Zeng, Valentin Deschaintre, Iliyan Georgiev, Yannick Hold-Geoffroy, Yiwei Hu, Fujun Luan, Ling-Qi Yan, and Miloš Hašan. Rgb-x: Image decomposition and synthesis using material- and lighting-aware diffusion models. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers '24*, SIGGRAPH '24, page 1–11. ACM, July 2024.

[53] Bo Zhang, Yuxuan Duan, Jun Lan, Yan Hong, Huijia Zhu, Weiqiang Wang, and Li Niu. Controlcom: Controllable image composition using diffusion model. *arXiv preprint arXiv:2308.10040*, 2023.

[54] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Scaling in-the-wild training for diffusion-based illumination harmonization and editing by imposing consistent light transport. In *The Thirteenth International Conference on Learning Representations*.

[55] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[56] Shuyang Zhang, Runze Liang, and Miao Wang. Shadowgan: Shadow synthesis for virtual objects with conditional adversarial networks. *Computational Visual Media*, 5:105–115, 2019.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We described the method and its quality.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discussed the limitation in the main content.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: We discussed the reason in the Exp section.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We described clearly in the method and exp section.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: We will release the code after acceptance.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We described clearly in the Exp section.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Don't need.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide them in the implementation section.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: No Broader impacts.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: No risk

    Guidelines:

    - The answer NA means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We properly cited them.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We described well the proposed dataset.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: We provide a user study in the supplementary material, and describe the method of investigation, sample capacity and quantitative results in detail.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This paper does not involve any crowdsourcing or research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.