Expressive-WICED
C. E. Catalano and A. D. Parakkat (Editors)

DepthLight: a Single Image Lighting Pipeline for Seamless Integration of Virtual Objects into Real Scenes

Raphael Manus[1,2], Marc Christie[1], Samuel Boivin[1], Pascal Guehl[2]

[1]Inria, IRISA, CNRS, Univ. Rennes, [2]LIX, Ecole Polytechnique/CNRS, IP Paris
https://depthlight.github.io

Figure 1: *A real image enriched with diffuse synthetic spheres rendered with different methods.* ***Left****: scene rendered using a regular image-based lighting technique (HDRi approach).* ***Right****: scene rendered using our DepthLight technique. Note how the regular HDRi approach fails to correctly capture the lighting, as the correct spatial properties of the light is not taken into account. In contrast, our technique better captures the light from the main window (see left, right and top spheres) and from the top light (see top sphere). In addition, it exploits an estimated 3D reconstruction of the scene to render coherent illumination from the scene on the sphere (see bottom left sphere).*

Abstract

We present DepthLight, a method to estimate spatial lighting for photorealistic Visual Effects (VFX) using a single image as input. Previous techniques rely either on estimated or captured light representations that fail to account for localized lighting effects, or use simplified lights that do not fully capture the complexity of the illumination process. DepthLight addresses these limitations by using a single LDR image with a limited field of view (LFOV) as an input to compute an emissive texture mesh around the image (a mesh which generates spatial lighting in the scene), producing a simple and lightweight 3D representation for photorealistic object relighting. First, an LDR panorama is generated around the input image using a photorealistic diffusion-based inpainting technique, conditioned on the input image. An LDR to HDR network then reconstructs the full HDR panorama, while an off-the-shelf depth estimation technique generates a mesh representation to finally build a 3D emissive mesh. This emissive mesh approximates the bidirectional light interactions between the scene and the virtual objects that is used to relight virtual objects placed in the scene. We also exploit this mesh to cast shadows from the virtual objects on the emissive mesh, and add these shadows to the original LDR image. This flexible pipeline can be easily integrated into different VFX production workflows. In our experiments, DepthLight shows that virtual objects are seamlessly integrated into real scenes with a visually plausible estimation of the lighting. We compared our results to the ground truth lighting using Unreal Engine, as well as to state-of-the-art approaches that use pure HDRi lighting techniques (see Figure 1). Finally, we validated our approach conducting a user evaluation over 52 participants as well as a comparison to existing techniques.

CCS Concepts

• ***Computing methodologies*** → ***Computational photography;*** *Image manipulation;* ***Mixed / augmented reality;*** *Computer graphics;*

DOI: 10.2312/exw.20251060

1. Introduction

Estimating the lighting properties of an existing real scene from partial information such as an image remains a complex and challenging problem due to its ill-posed nature. This is pertained to (i) the intrinsic complexity of scene lighting (ambient lighting, direct and indirect illumination, reflections, shadows, specular highlights) together with the underlying difficulty in untangling light contributions without prior information, (ii) the lack of knowledge of the 3D scene geometry that guides the light interactions, and (iii) the uncertainty in the positions and characteristics of the light sources in the scene, as well as on the object materials. While many techniques have relied on light probes [Deb98] and template objects placed in the real scene [BM14; YAH*23] to improve the estimations, these approaches remain complex to set up and configure.

Recent trends have explored the use of single-view images as input for illumination estimation in indoor and outdoor scenes [DEHL23; LMF*19]. The challenge is further complicated when the input image has a low dynamic range (LDR) and a limited field of view (LFOV). Such approaches either estimate a single HDRi map from the image (replicating the use of reflective diffuse and glossy balls to capture incoming light characteristics [PCS*24]), or estimate positions and characteristics of light sources [HW12; GHS*19; LMF*19]. In the first case, the techniques lack spatial characteristics of lighting since HDRi maps act as infinite light sources. In the second case, the techniques do account for local characteristics but do not model the scene geometry which strongly influences the lighting process.

In this paper, we introduce DepthLight, a novel pipeline for estimating spatial lighting as an extension of Debevec's light-based model [Deb98]. In contrast with a light-based model, DepthLight generates an estimated 3D mesh around the input image from a depth map estimation, and projects an estimated HDRi map on this mesh to create an *emissive textured mesh* which illuminates the synthetic objects and casts coherent shadows on the image.

Generating such an emissive mesh requires to (i) know the 3D geometry of the scene around the image, and (ii) have HDR representation of the scene illumination. The approach we propose in this paper is to first extend the input LFOV image to a 360 panorama using PanoDiff, a diffusion-based inpainting technique [WCL*23] which provides a probabilistic estimation of the surrounding scene, conditionned by the LFOV image. This LDR panorama image is then converted to an HDR panorama image using PanoLANet, an off-the-shelf LDR to HDR technique [YLL*21]. A depth estimation technique (DepthAnythingV2 [YKH*24]) is then applied on the LDR panorama, from which a 3D mesh can be constructed. This mesh provides a rough, yet valuable, estimation of the surrounding scene around the LFOV image. The HDR panorama image is then applied as an emissive texture on the 3D mesh, and exploited to (i) render the virtual objects using the emissive HDRi mesh as light sources, and (ii) render the shadows of the virtual objects on the 3D emissive mesh. Both renderings are then composited with the original image.

Unlike multi-view inverse rendering methods, our single-view approach allows seamless integration into production pipelines without any additional camera or point of view. This makes our solution both efficient and straightforward, avoiding the complexities of volumetric light representations and custom rendering engines, and ensuring compatibility with existing VFX workflows.

With only a single-view image as input, our approach does not produce a precise estimation of the light sources, but rather a plausible estimation providing visually convincing results. This focus aligns with industry needs, where visual integration takes precedence over precise lighting when lacking full information on the scene. In addition, the technique offers a flexible, multi-input pipeline, accommodating various input stages, such as for example an LDR LFOV image, a 360° LDR image directly, or a HDR panorama. We validate our approach using a user evaluation where participants are asked to evaluate the visual quality of our technique (see section 4.5), as well as against classical light estimation techniques. We also provide a complete comparison with existing methods using the Lightsome framework (see section 4.4).

In summary, the main contributions of our method are:

- a complete and fully automatic pipeline for the seamless insertion of virtual objects into real scenes from a single image;
- an estimation of the spatially varying lighting using an emissive mesh in replacement of traditionnal HDR maps;
- a rendering pipeline to illuminate the virtual objects using the emissive mesh, and integrate the casted shadows of virtual objects on the original image by exploiting the 3D mesh estimation.

2. Related Work

2.1. Lighting Estimation

Lighting estimation is critical for various applications, such as augmented reality and VFX, for example. It has a rich history and is a very prolific topic in both computer vision and computer graphics [JS20]. Many approaches use external sources of information such as light probes [Deb98] or implicit sensors that use arbitrary objects in the scene [BM14; YAH*23]. Some works focus on multi-view images to reconstruct the scene using inverse rendering techniques [LHL*24; PMGD21] and intrinsic decomposition [LLYL20].

We decided to focus on a robust sensor-less approach using only a single LDR LFOV image. A number of methods exist for such cases but often work only either for indoor situations [GHS*19; GSH*19; WGL22; ZZY*21], outdoor situations [HSH*17; HAL19; ZSH*19], or both [DEHL23; LMF*19]. Most of those techniques rely on predicting an HDR environment map, either explicitly as an HDR texture [PCS*24] or implicitly using a parametric lighting representation [HW12; GHS*19; LMF*19]. A lot of these methods use neural networks to predict their lighting estimation [GSY*17; WYLL22; DEHL23].

However, a single environment map used as an infinite light source is not enough to correctly describe spatial lighting effects. Therefore, many works estimate spatially-varying lighting in different manners. Some approaches assume the geometry is known through multi-view depth images [BM13; MKC*17] or multi-view scene reconstruction [ZCC16; LHL*24], or require manual annotation [KHFH11], stereo pairs of images [SMT*20], or known depth [ZYZ*22]. Our approach does not require any prior knowledge.

Some work estimates a separate environment map for each pixel

in the input image [GSH*19; LYO*23; BHY*23] which is not a lighting representation commonly used by artists. Others infer 3D light locations and parameters [GHS*19] or they use a combination of 3D light and panorama [WGL22; SJT*24]. Typically VISPI [SJT*24] optimizes up to three light source positions from depth and albedo estimations of panoramic views, yet only consider direct light sources and lack shadows. A few approaches use complex volumetric representations of lighting [SMT*20; WPFK21], which are not user-friendly and are not meant to be used in consumer software. A combination of HDR skydomes and volumetric lighting has been proposed in [WCA*22]. Spatial estimates of radiances are performed through scene discretization and results are enhanced by regressing light estimates using realistic images through a differential renderer and adversarial training. Some GANs and diffusion-based techniques focus on outpainting an input image to a 360° panorama [DHE*22; AMA22; HMH22; WCL*23], but perform poorly in light estimation due to their LDR prediction. Our goal is to obtain a user-friendly representation that can be used in most rendering software natively, especially for the VFX industry.

Our approach uses PanoDiff [WCL*23], a latent diffusion model for LDR panorama generation, allowing to predict high-quality, realistic LDR environment maps from a single LDR POV image.

2.2. HDR Reconstruction

HDR imaging has been used for a long time to recover radiance information from a scene [DM97]. This is because HDR maps can be used as the illumination of the scene to approximate a full reflectance model of the scene [Deb98]. However, proper HDR imaging requires an appropriate setup and additional tools, a global process that can be time-consuming. This is why there are many proposed solutions to extrapolate HDR information from a single LDR input image using machine learning techniques, but most of these methods [WSP*23][MBHD18] focus on HDR reconstruction for display applications such as HDR monitors [AFR*07]. Some works focus on image-based lighting (IBL) applications but they are often restrictive, limiting their use to only outdoor scenes [ZL17] or only sky images [SYMH23]. We incorporate LANet [YLL*21] in our pipeline, which is a HDR reconstruction network designed for IBL trained on both indoor and outdoor HDR panoramas.

A number of direct HDR panorama reconstruction methods like StyleLight [WYLL22] and DiffusionLight [PCS*24] have been proposed. Such approaches are less focused on achieving photo-realistic results compared to LDR panorama reconstruction solutions. They indeed perform worse on estimating coherent geometric features of the scene, which are required for better depth estimation, and they are less appropriate for high-quality reflections casted onto inserted virtual objects. In contrast, our two-steps approach, which uses LDR panorama estimation [WCL*23] followed by HDR reconstruction [YLL*21], enables the creation of realistic HDR panoramas, which coupled with depth estimation [YKH*24], represents our spatial lighting representation.

3. Proposed Method

Given a LDR LFOV input image, our goal is to estimate the scene's illumination as a spatial representation, to render and relight a synthetic object, and cast its shadows on the input image. Our work is based on an emissive mesh estimation, followed by two rendering passes. These are illustrated in our unified pipeline for diffuse lighting estimation (see Fig. 2 for an overview of the pipeline).

Similarly to [Deb98], a key simplification is to consider the scene's materials as only being purely diffuse (specular surfaces being a possible enhancement of our work through albedo estimation techniques, *e.g.* [LLYL20]. In our process, we start with panorama estimation (see 3.2), then HDR reconstruction (see 3.3), a depth estimation (see 3.4) from which we reconstruct the emissive mesh, before performing the rendering and compositing passes. Necessary background elements are detailed in the next section (see 3.1).

Our pipeline can also seamlessly integrate input data from a 360° camera. In such cases, there is no need to extrapolate the environment map, and the input can be directly plugged into the subsequent stages of the pipeline for HDR reconstruction.

3.1. Background Knowledge

When facing the problem of integrating a synthetic object into an LDR LFOV image, we need to estimate the complex light interactions between the scene and the newly placed object, as well as the materials of the scene. A common simplification is to consider that the whole real scene, denoted r, consists in diffuse surfaces [Deb98]. Henceforth, we can assume that behind each pixel of the input image, there is a planar patch surface which follows the expression of the discrete radiosity equation as:

$$B_i^r = E_i^r + \rho_i^r \sum_{j=1}^{n} F_{ij} B_j^r \tag{1}$$

where we consider that each unit of surface has a constant radiosity B_i^r and constant reflectivity ρ_i^r, and F_{ij} represents the form factor for the radiation leaving patch j and hitting patch i. E_i^r represents the emittance of the surface. We will assume a diffuse-only hypothesis, and show how this assumption simplifies the lighting estimation and rendering process.

3.2. Panorama Reconstruction

For the panorama reconstruction, we use PanoDiff [WCL*23], a method that generates LDR panorama from a LFOV image using a novel latent diffusion-based model. A particular strength of the method is the ability to generate consistent high-quality and realistic content. PanoDiff uses custom ControlNet [ZRA23] units to exploit the input image as a control signal for a pre-trained Stable Diffusion [RBL*22] model that they train on panoramas. The input is our LDR LFOV image I, the field-of-view fov and the tilt information θ of the input image which goes through an equirectangular projection operation $\mathcal{F}_{\text{proj}}$, and produces a partial panorama image P and a visibility mask M, formulated as :

$$\mathcal{F}_{\text{proj}}(I, fov, \theta) \rightarrow P, M$$

The latent diffusion model (LDM) from PanoDiff is formulated as a function $\mathcal{F}_{\text{LDM}}$ taking the previous partial panorama image P and visibility mask M to generate an image $LdrPano$. The projection function $\mathcal{F}_{\text{proj}}$ takes the input image, horizontal fov and tilt

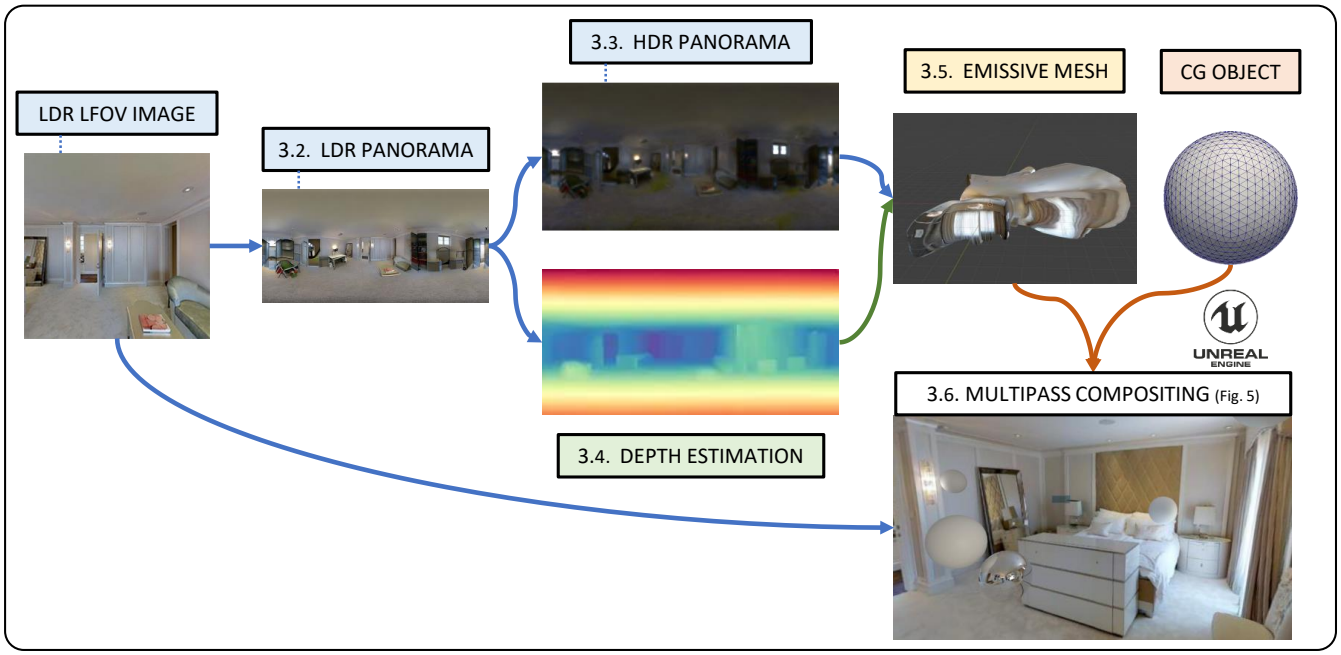

Figure 2: *Overview of our DepthLight pipeline. Blue dashed lines represent the possible input points of our pipeline. Our testing involved the whole pipeline using LDR LFOV input images. Numbers 3.2, 3.3, 3.4, 3.5 and 3.6 refer to dedicated sections in the article.*

angle of the input image and projects it in equirectangular projection [Hei00] in a partial panorama and a visibility mask. This mask indicates to the LDM the parts to outpaint (see Fig. 3).

While PanoDiff utilizes a custom pretrained Stable Diffusion [RBL*22] model for this task, we also explored the use of Stable Diffusion with a LoRa [HSW*21] trained on equirectangular panoramas and ControlNet [ZRA23] for image inpainting. This approach generates results much faster than PanoDiff, although it lacks comparative metrics to Ground Truth (GT).

3.3. HDR Reconstruction as a Light Source Radiance Approximation

The panorama generated from the PanoDiff LDM contains realistic and high-quality LDR content. However, it cannot be used as radiance information because of the fundamental differences between LDR and HDR representations [DM97]. HDR imaging has been used as a representation of a scene radiance [DM97; Deb98] as it represents both low levels of indirect radiance from surfaces, and high levels of direct radiance from light sources, to easily simulate the effects of indirect illumination from the environment. To address this, we use a luminance attentive network, called PanoLANet [YLL*21], for HDR panorama reconstruction from an LDR panorama (see Fig. 4).

PanoLANet differentiates HDR reconstruction between display and rendering applications, the latter requiring precise luminance for Image-Based Lighting (IBL). This distinction is crucial because for display applications HDR reconstruction mostly aims to create visually pleasing results, whereas for rendering applications it requires accurate light measurements.

Figure 3: ***Left****: LFOV input image.* ***Middle****: equirectangular projection of input (fov is 72°) and visibility mask.* ***Right****: LDR panorama reconstruction using PanoDiff [WCL*23] diffusion model for photorealistic panorama outpainting.*

Figure 4: *HDR panorama reconstruction using PanoLANet [YLL*21].* ***Left****: input panorama at 0 EV (normal exposure).* ***Right****: PanoLANet's reconstruction, retaining more light information at -5 EV. This allows our technique to extract the relevant radiance information for our emissive mesh.*

3.4. Depth Estimation as a Form Factor Approximation

To achieve a spatially coherent lighting, we need to determine the 3D positions of the surrounding materials and lights which irradiate the scene. Given the limited knowledge from the original image, we propose to build on an approximate mesh representation of the scene around the image, which can be built by using off-the-shelf depth estimation techniques [YKH*24] applied to our SDR generated panorama. DepthAnything v2 is a robust monocular depth estimation model. Although not specifically trained on 360° equirect-

Figure 5: *Depth estimation comparison for a fully synthetic scene.* ***Top****: ArchViz Interior scene from Unreal Engine's marketplace.* ***Middle****: Ground truth depth map associated with the scene.* ***Bottom****: Depth map estimation of the scene [YKH*24].*

angular panoramas, its generalization ability – rooted in large-scale training on diverse real-world data – enables plausible depth estimation in such settings. An additional quantitative evaluation could be performed on rectangular patches of the panorama to properly assess the generalization capacity, yet results produced provided visually coherent depth estimations (see Fig. 5 for a comparison). Not only does this depthmap computation provide an estimate of the distances from the camera to the surrounding scene, but it also provides an approximation of the form factor in the global illumination equation (see section 3.1); such an approximation is given by the ability to query mutual visibility and distances between different regions of the image, which we can exploit to relight synthetic objects (see Section 3.5).

We formulate the DepthAnything output as a *Disparity* map generated by a function $\mathcal{F}_{\text{DA}}$ from the input image *LdrPano*. We can express the discrete radiosity equation of the virtual object s as:

$$B_i^s = E_i^s + \rho_i^s \left(\sum_{k=1}^{m} F_{ik} B_k^s + \sum_{j=1}^{n} F_{ij} B_j^r \right) \tag{2}$$

where B_i^s represents the radiance of a surface unit i of the virtual object s given its own emittance E_i^s and reflectivity ρ_i^s. The object's radiance is then influenced by its own radiance (B_k^s) (self illumination) and also by the real scene radiances (B_j^r).

The form factor equation F_{ij} is classically expressed as the ratio of energy leaving a unit of surface j and arriving to the surface i as a function of their distances, occlusions and angles:

$$F_{ij} = \frac{\cos\theta_i \cos\theta_j A_j}{\pi r^2} V_{ij} \tag{3}$$

where A_j is the area of the unit of surface j, r is the distance between surface i and j and $\cos\theta_i, \cos\theta_j$ are the respective angles of surfaces i and j, where the visibility term V_{ij} is either 1 or 0, whether the surface i sees the surface j or not, respectively. By exploiting the estimated depth information, we can provide a reconstructed mesh from the scene, as a placeholder of the real scene geometry. That mesh can then be used to estimate the V_{ij} term of equation 3, by providing more precise information than using spherical HDRI maps or box-shaped representations [DM97].

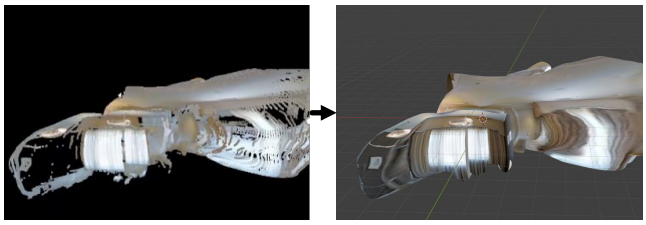

Figure 6: *Left: Point cloud of a scene reconstructed from the depth estimation. Right: Mesh reconstruction of the same scene.*

3.5. Emissive Mesh as an Approximation of the Global Illumination Equation

We define an emissive mesh as being the combination of a 3D geometrical mesh and its associated diffuse textures considered as radiances for illuminating objects in the scene. We therefore reproject the real image onto the reconstructed mesh, which we can then consider as an emissive mesh, *i.e.* for all faces of this mesh, we consider $B_j^r = \xi_j^r$ where ξ_j^r is the estimated emissive radiance. Note that, compared to an HDRI projected to a spherical surface, the form factor is no longer estimated, as it integrates the distance to faces, the occlusions and the relative angles. The only approximation that we do here is related to the capacity of the depth estimator to properly predict the relative scene depth.

Equirectangular depth maps can be mapped in 3D to visualize them as a 3D point cloud. This is done by creating a uniform grid of points on the unit sphere of the dimension of the depth map and then scaling each vector of the grid by its corresponding depth value. This is especially useful in our case as it allows us to estimate the 3D scene of the panorama.

The point cloud can then be triangulated since the point ordering is derived from the pixel ordering of the depth map. Each point is considered as a vertex of the mesh and is stored in a flattened array.

We then map the HDR environment onto the mesh which can be integrated into any renderer capable of rendering emissive meshes. Figure 6 shows the difference between the point cloud representation and scene's mesh with the reprojected HDR map as a texture.

3.6. Real-Time Compositing in Unreal Engine

Since our goal is a seamless integration of a virtual object in a real scene, we have to perform that integration either directly in a renderer or afterwards. Our approach to compositing is similar to the differential rendering technique as described by [Deb98].

We chose Unreal Engine (UE) for its powerful real-time photorealistic rendering capabilities, enabling the real-time compositing of relighted virtual objects in our scenes.

The impact of the synthetic object lighting over the real scene is done through the compositing pipeline and the photorealistic UE renderer. If we refer to section 3.5, this means that equation 2 can be expressed as a sum of the estimated emissive radiances ξ_i^r and the radiance from the virtual object faces.

$$B_i^r = \xi_i^r + \rho_i^r \sum_{k=1}^{m} F_{ik} B_k^s \tag{4}$$

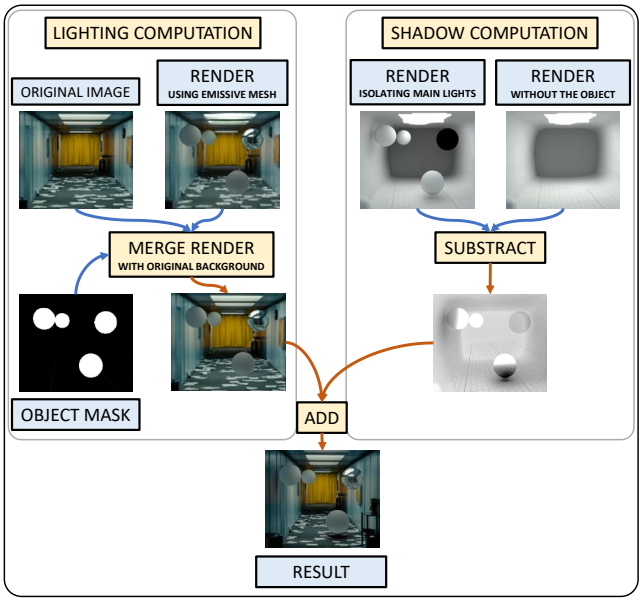

Figure 7: *Our compositing pipeline for Unreal Engine. We create a two-pass compositing pipeline for shadow casting and lighting as they operate completely differently. The shadow pass allows to have any synthetic object casting shadows over the real scene. The lighting pass allows to correctly illuminate any synthetic object using our spatially-varying emissive mesh.*

Given the hypothesis on the diffuse nature of the scene surfaces, the reflectivity ρ_i^r is Lambertian, represented as a constant.

To achieve the rendering and compositing in UE, we used two main features: the emissive material of objects and the mesh distance fields of UE's Lumen renderer. Instead of using the classical spherical environment map (skylight) combined with an HDR, we linked the emissive material property of the real scene 3D mesh to the HDR image, both coming from our pipeline. This approach is typically used to simulate the emissive radiances ξ_i^r of equation 4.

Moreover, a mesh distance field in UE refers to a volume representation of the distance to the nearest surface of the mesh, allowing for efficient calculations for effects like ambient occlusion and emissive lighting. It is thus used here to compute the form factor F_{ik} of equation 4. The remaining terms of equation 4 are computed by the global illumination renderer of UE (Lumen). Note that in terms of implementation details we had to add a thickness to our mesh and flip the interior normals, in order to have the mesh distance fields and the global illumination to work properly.

In addition to that setup within the engine, we have created a specific pipeline to handle the final compositing and the shadow casting of the inserted virtual elements within the input image. That compositing pipeline in UE is detailed in Fig. 7.

4. Results

In Table 1, we compare our DeepLight technique to the existing related works. The *spatially varying light* feature corresponds to whether the work accounts for infinite light sources (as with environment maps) or localized lights. The *Comparison to CG scene as GT* is related to whether the work uses a complete synthetic rendering to compare the estimated results with GT rendered results. The *comparison to HDR GT* feature characterizes the work which compares their estimated lighting with a ground truth lighting obtained with an HDR probe device. The *Complete compositing pipeline* specifies whether the technique is capable of rendering the lighting and the shadows of the synthetic object in the real scene and also handling occlusions between the real and the CG objects. The *360° panorama support* feature characterizes whether the technique can use a 360° panorama as input. Finally, the *3D scene model estimation* feature mentions whether the technique reconstructs at least partially the real scene, which typically enables occlusions of synthetic objects by real ones.

	Spatially varying light	Comparison to CG scene as GT	Comparison to HDR GT	Complete compositing pipeline	360° Panorama support	3D scene model estimation
Weber [WGL22]		✓	✓			✓
Zhang [ZSH*19]		✓	✓		✓	
EverLight [DEHL23]	✓		✓			
StyleLight [WYLL22]			✓		✓	
Gardner [GHS*19]	✓	✓	✓		✓	
Khan [KRFB06]	✓	✓	✓			
DepthLight (ours)	✓	✓	✓	✓	✓	✓

Table 1: *Feature Comparison between existing single image techniques and our Depthlight.*

Furthermore, we evaluate the performance of our method in both indoor and outdoor scenarios (even if our method is not initially designed for outdoor scenes) and make comparisons with SOTA methods. There is no standardized way for evaluating light estimation. Moreover, quantitative metrics do not necessarily reflect the actual human perception of realism and may not match with what a human might believe be a photorealistic and/or plausible result [GDH*24]. The controlled psychophysical study [GDH*24] reveals that humans' preference contradicts usual image metrics in the vast majority of cases. Instead, they propose a framework called Lightsome for standardized evaluation of lighting estimation techniques, as well as a novel comparison metric to evaluate the perceptual preference between two renders. We use that metric in the next section to compare our results to existing techniques. We have also conducted a user evaluation over 52 participants to validate the quality and the visual plausability of our technique (see 4.5).

4.1. Implementation Details

Using the Lightsome framework [GDH*24], we rendered a virtual sphere with either diffuse or glossy material under different lighting estimations for each evaluated model. Though our technique is not

designed to handle outdoor scenes, we maintained both indoor and outdoor scenes in the evaluation for comparison.

We compared our approach to both non-spatial representation models: StyleLight [WYLL22] for indoor, [ZSH*19] for outdoors, EverLight [DEHL23] for both; and indoor spatial representation models [WGL22; GHS*19]. We also compared these methods to a simpler approach proposed by [KRFB06], which consists of projecting the input image onto a sphere and then mirroring it to create an LDR environment map without extrapolating information from the input image. These scenes were chosen to cover a range of illumination scenarios, including indoor and outdoor scenes, variations in light color temperature, intensity, and directional properties.

4.2. Performances

In the DepthLight pipeline, PanoDiff requires 5s per image. LDR to HDR and DepthAnythingV2 [YKH*24] are only inferences, and take an average of 12ms (on a GTX4090). The mesh is reconstructed instantly and is rendered in real-time at 60fps in UE.

4.3. Quantitative Evaluation

The Lightsome evaluation framework [GDH*24] does not take into account the light's spatial variations, and neither do EverLight [DEHL23], StyleLight [WYLL22], or Khan [KRFB06]. For comparison, we simply render the scenes with our emissive mesh centered on the camera position. We evaluate the accuracy of each method by computing the VIF metric [GDH*24] and the BRISQUE metric [MMB12], a referenceless IQA supposed to evaluate "naturalness" of an image, as proposed by [GDH*24]. Results for each metric for indoor and outdoor evaluations are shown Table 2. For each metric, D is the evaluation rendered with a Diffuse material and G is for a Glossy material.

Results show that our model performs either on par and often better than current techniques. In the cases where our model is not the best one, the measured error is always close to the best result. Moreover, it is fair to consider that removing the spatial variations aspect of our technique by using such a simple scene (a sphere) for the evaluation is not necessarily the best way to compare it to existing techniques which completely omit that part.

4.4. Qualitative Comparison

We present qualitative comparisons to other models in Table 10. We show the lighting estimation of each model being represented as an environment map in the last row. Our method provides both plausible lighting and reflections for both indoors and outdoors. However, we would like to point out that using an isolated sphere + an HDR image as ground truth is not appropriate for DepthLight (or any similar technique), as the scene is not designed for spatial variations of the light. That comparison method is also highly biased by the panorama generator technique (see the huge differences between light source estimations for example in Fig. 10).

We also compared our technique to a regular HDRI approach, as defined in [DM97; Deb98], using the same estimated HDR map but through our emissive mesh representation (see figures 8 and 1). We can see that a regular HDRI approach fails to properly simulate the lighting as there is no spatial awareness of the scene.

Finally, We utilize a scene from Unreal Engine's Marketplace to compare our DepthLight to the actual ground truth. Here, a fully synthetic scene is photo-realistically rendered in UE using their renderer Lumen. We then use a single image rendered from a static point of view of that scene as the only input to our DepthLight pipeline. Finally, we render an armadillo model in both UE (which is considered here as the ground truth) and compare it to the same armadillo rendered using our DepthLight (see Fig. 11).

4.5. User Evaluation

We evaluated the visual integration quality of our method by providing 2 images for 10 different scenes having inserted CG objects to 52 participants (19 women and 33 men), aged 21 to 56 years old. A typical scene example is shown in Fig. 11. Participants were also made aware that one image was generated using a state-of-the-art rendering software (considered as the real image), while the other one was fully generated using our method (considered as the fake image). Users had to objectively evaluate both images following 4 criteria (3 using the Likert scale from 1 (bad) to 7 (perfect)): visual coherency of the lighting, shadow realism , quality of insertion for the added object (armadillo), and finally A/B testing asking them to decide which image was real and which one was fake. The results are shown in Table 3 as the nearest rounded average integers, and demonstrate that the users were unable to decide which image was real and which was fake, as 53% of the participants chose our image as the real. This leads us to consider that DepthLight is capable of rendering visually plausible scenes.

5. Limitations and Future Works

Our current method uses a simplified lighting model that does not take into account the BRDFs of objects in the scene. This means that we do not account for how light reflections of a virtual object cast light on a real one (color bleeding). A possible extension could consist in using a fully automatic estimate of BRDFs directly from the input image, such as [LLYL20], to better estimate how radiances emitted from the synthetic object can affect the real image. This requires to extend our compositing pipeline with an additional pass allowing for our synthetic objects to also relight the real scene. Additionally, our compositing pipeline (see Fig. 7) does not work iteratively, meaning that real objects do not directly cast shadows over synthetic objects. To improve the approach we should also better estimate the behavior of depth estimation on panorama views, for example through image tiling techniques.

Another area of improvement is the use of modern *differentiable rendering techniques* [JSRV22]. These techniques could help refine our lighting models by comparing our estimated scene with the actual scene, making the results more accurate and realistic, or exploiting realistic image features to regress the lighting parameters using adversarial training as in [WCA*22].

As described in section 4, our method is not designed for outdoor scenes, especially because DephtAnythingV2 fails to correctly estimate the position of the sun. An interesting future work could ex-

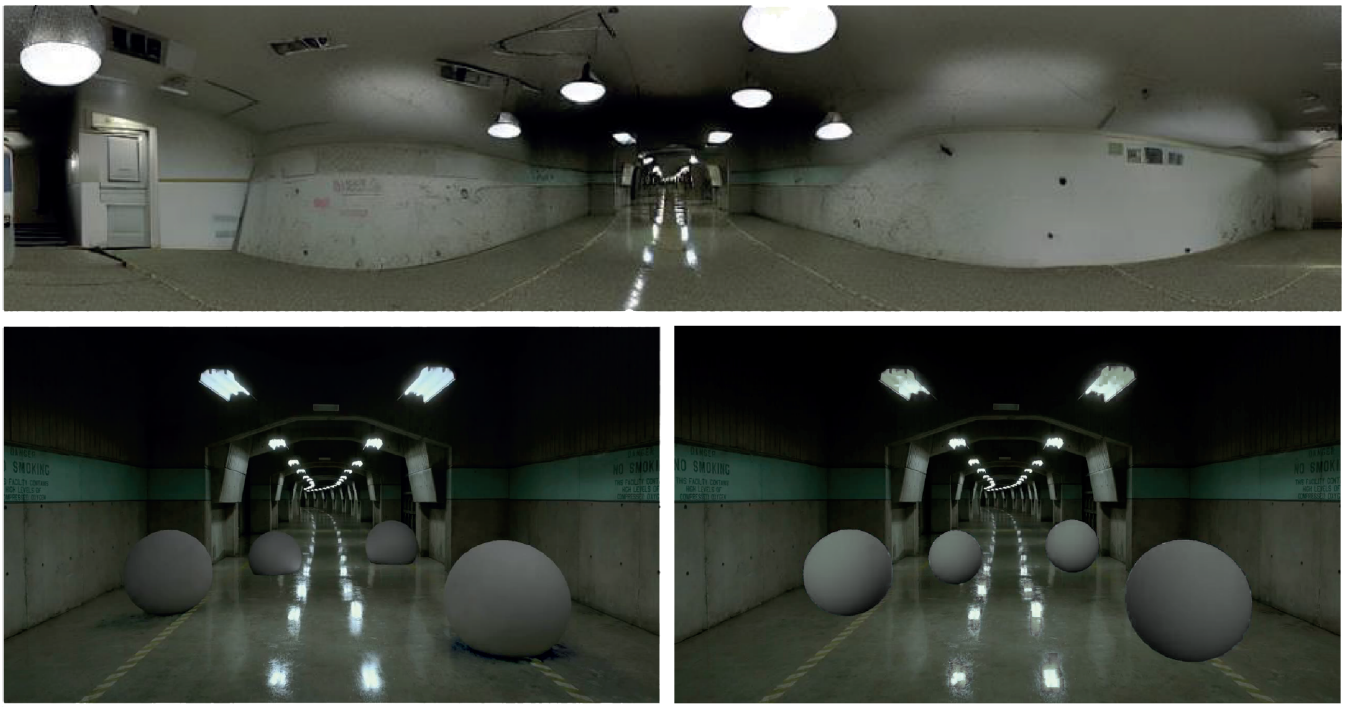

Figure 8: ***Top****: Estimated HDR panorama from the bottom real scene with four virtual spheres.* ***Bottom left:*** *Our DepthLight technique.* ***Bottom right:*** *Regular HDRI relighting technique. Using DephtLight, note the quality of the lighting interactions and the correct orientation of the simulated shadows casted by the virtual objects onto the real scene. Regular HDRI techniques fail to correctly simulate the final lighting and would need many manual adjustments to be realistic enough for VFX as they do not consider the depth in the scene. This may also generate optical illusions as the spheres in the right image are not intersecting with the floor because there are no depth information. Note also that the lighting of the spheres in the HDRI result tends to generate overlit objects as there are no distance-to-light factor taken into account, as opposed to our emissive mesh method.*

Figure 9: *Two qualitative examples showcasing DepthLight's capability to generate spatial-aware lighting effects from a single LDR and LFOV image. We integrated two 3D rabbits, half diffuse and half specular, in each input image, taken from the [GSH*19] dataset. DepthLight reconstructs an entire emissive texture mesh of the scene that re-casts spatialized lighting on virtual objects and enables seamless integration of synthetic objects in real images, whatever their location in the scene. This approach enables photorealistic relighting of objects, capturing intricate lighting variations and occlusion effects that traditional environment map representations fail to achieve.*

ploit a more evolved depth estimation technique for outdoor scenarios, through hybrid representations [WCA*22]. Currently, DepthLight is designed for single images and does not handle dynamic lighting in video sequences. For videos with a limited field-of-view (LFOV), a possibility is to compute lighting for one frame and apply it to subsequent frames using 3D camera tracking. Future

	Indoor				Outdoor			
	VIF↑		BRISQUE↓		VIF↑		BRISQUE↓	
	D	G	D	G	D	G	D	G
Weber [WGL22]	0.918	0.749	62.124	58.761	–	–	–	–
EverLight [DEHL23]	0.684	0.744	61.101	54.798	0.761	0.693	55.585	51.242
StyleLight [WYLL22]	0.883	0.707	61.878	59.327	–	–	–	–
Gardner [GHS*19]	0.822	0.695	62.693	60.736	–	–	–	–
Khan [KRFB06]	0.693	0.698	62.640	56.686	0.744	0.642	55.402	51.145
Zhang [ZSH*19]	–	–	–	–	1.018	0.774	56.076	52.945
DepthLight (ours)	0.879	0.762	60.317	50.680	0.901	0.752	56.459	50.672

Table 2: *Lightsome framework [GDH*24] for D diffuse and G glossy materials from fig. 10. Even if our method is designed to handle Indoor scenes, we added comparisons with Outdoor scenes, that shows our results are comparable, or close to state-of-the-art techniques. The table includes VIF↑ and BRISQUE↓ metrics for comparison across different methods. We highlight best values and second best. With most metrics/materials, our method performs better than state-of-the-art in Indoor scenes (with the exception of Diffuse on the VIF metric).*

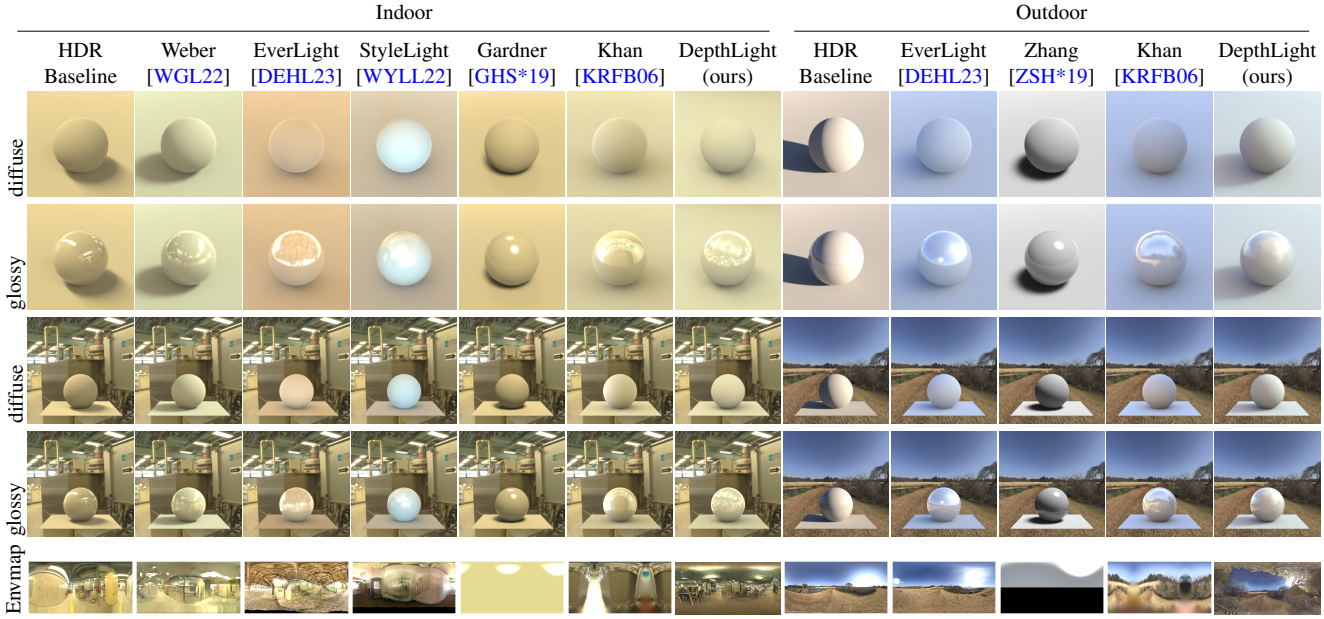

Figure 10: *Qualitative comparison using the Lightsome evaluation framework [GDH*24] of both indoor and outdoor scenarios. The leftmost column of both scenarios shows the baseline HDR environment map used for lighting. For a fair comparison with other models, we ignore spatial variations in this experiment. We display both the rendered sphere without the input as background and the sphere from further away with the input as background.*

research in generative AI could help create temporally consistent lighting for LFOV video inputs.

6. Conclusion

DepthLight provides a complete pipeline for seamlessly integrating virtual objects into real scenes, using a single LFO image. It addresses the limitations of traditional HDRI-based techniques by using a novel emissive texture mesh representation, thus allowing for spatial and occlusion lighting effects. Experimental results demonstrate its effectiveness and potential, and we believe it is an interesting step forward to enhance and ease VFX and AR workflows. We validated the technique by not only comparing our results to a ground truth representation using Unreal Engine but also conducting a user evaluation based on the visual quality of our approach.

References

[AFR*07] AKYÜZ, AHMET OĞUZ, FLEMING, ROLAND, RIECKE, BERNHARD E, et al. "Do HDR displays support LDR content? A psychophysical evaluation". *ACM Transactions on Graphics (TOG)* 26.3 (2007) 3.

[AMA22] AKIMOTO, NAOFUMI, MATSUO, YUHI, and AOKI, YOSHIMITSU. "Diverse plausible 360-degree image outpainting for efficient 3dcg background creation". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, 11441–11450 3.

	Unreal Engine			DepthLight (ours)		
	1-3	4-5	6-7	1-3	4-5	6-7
Coherent Lighting	8	39	5	6	26	20
Realistic Shadows	15	35	2	0	40	12
Overall Quality of armadillo's insertion	10	40	2	3	30	19

	Unreal Engine		DepthLight (ours)	
	Fake	Real	Fake	Real
A/B Testing	28 (53.8%)	24 (46.2%)	24 (46.2%)	28 (53.8%)

Table 3: *User Evaluation of DepthLight over 52 participants. The **coherent lighting** refers to the quality of the global lighting (global illumination in the scene including the armadillo). The **shadow realism** only concerns the shadows generated by the armadillo. The **overall quality** is to verify the final result including the consistency in user's answers. Finally, during the **A/B testing**, people were asked to decide which image was generated using our method and which one was the actual reference ground truth image, among 10 different scenes. It is interesting to note that users have often chosen the quality of our object insertion over Unreal Engine, even for the armadillo's shadows. Note that for the sake of clarity, all average numbers have been rounded to the nearest integer.*

Figure 11: *Comparison between Ground Truth and DepthLight.* ***Top****: Original interior scene from Epic Games.* ***Bottom Left****: UE rendering of an inserted armadillo inside the original scene from a fixed point of view, used as ground truth here.* ***Bottom Right****: Our DepthLight result by inserting the same armadillo inside the rendered image (without the armadillo) but using our pipeline. Using a single image of the original scene, the DepthLight pipeline estimates the lighting to render the armadillo into the original image. The lighting difference comes mostly from the panorama generation technique. Note however that the result remains of high quality and it is hard to tell which image actually yields the most visually acceptable result.*

[BHY*23] BAI, JIAYANG, HE, ZHEN, YANG, SHAN, et al. "Local-to-Global Panorama Inpainting for Locale-Aware Indoor Lighting Prediction". *IEEE Transactions on Visualization and Computer Graphics* (2023) 3.

[BM13] BARRON, JONATHAN T and MALIK, JITENDRA. "Intrinsic scene properties from a single rgb-d image". *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2013, 17–24 2.

[BM14] BARRON, JONATHAN T and MALIK, JITENDRA. "Shape, illumination, and reflectance from shading". *IEEE transactions on pattern analysis and machine intelligence* 37.8 (2014), 1670–1687 2.

[Deb98] DEBEVEC, PAUL. "Rendering synthetic objects into real scenes: Bridging traditional and image-based graphics with global illumination and high dynamic range photography". *Acm siggraph 2008 classes*. 1998, 1–10 2–5, 7.

[DEHL23] DASTJERDI, MOHAMMAD REZA KARIMI, EISENMANN, JONATHAN, HOLD-GEOFFROY, YANNICK, and LALONDE, JEAN-FRANÇOIS. "EverLight: Indoor-outdoor editable HDR lighting estimation". *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, 7420–7429 2, 6, 7, 9.

[DHE*22] DASTJERDI, MOHAMMAD REZA KARIMI, HOLD-GEOFFROY, YANNICK, EISENMANN, JONATHAN, et al. "Guided co-modulated GAN for 360 field of view extrapolation". *2022 International Conference on 3D Vision (3DV)*. IEEE. 2022, 475–485 3.

[DM97] DEBEVEC, PAUL E and MALIK, JITENDRA. "Recovering high dynamic range radiance maps from photographs". *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*. 1997, 643–652 3–5, 7.

[GDH*24] GIROUX, JUSTINE, DASTJERDI, MOHAMMAD REZA KARIMI, HOLD-GEOFFROY, YANNICK, et al. "Towards a Perceptual Evaluation Framework for Lighting Estimation". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, 4410–4419 6, 7, 9.

[GHS*19] GARDNER, MARC-ANDRÉ, HOLD-GEOFFROY, YANNICK, SUNKAVALLI, KALYAN, et al. "Deep parametric indoor lighting estimation". *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, 7175–7183 2, 3, 6, 7, 9.

[GSH*19] GARON, MATHIEU, SUNKAVALLI, KALYAN, HADAP, SUNIL, et al. "Fast spatially-varying indoor lighting estimation". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, 6908–6917 2, 3, 8.

[GSY*17] GARDNER, MARC-ANDRÉ, SUNKAVALLI, KALYAN, YUMER, ERSIN, et al. "Learning to predict indoor illumination from a single image". *arXiv preprint arXiv:1704.00090* (2017) 2.

[HAL19] HOLD-GEOFFROY, YANNICK, ATHAWALE, AKSHAYA, and LALONDE, JEAN-FRANÇOIS. "Deep sky modeling for single image outdoor lighting estimation". *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, 6927–6935 2.

[Hei00] HEIDRICH, WOLFGANG. "Environment Maps and Their Applications". *Max-Planck-Institute for Computer Science, Saarbrücken, Germany* 19 (2000) 4.

[HMH22] HARA, TAKAYUKI, MUKUTA, YUSUKE, and HARADA, TATSUYA. "Spherical image generation from a few normal-field-of-view images by considering scene symmetry". *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.5 (2022), 6339–6353 3.

[HSH*17] HOLD-GEOFFROY, YANNICK, SUNKAVALLI, KALYAN, HADAP, SUNIL, et al. "Deep outdoor illumination estimation". *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, 7312–7321 2.

[HSW*21] HU, EDWARD J, SHEN, YELONG, WALLIS, PHILLIP, et al. "Lora: Low-rank adaptation of large language models". *arXiv preprint arXiv:2106.09685* (2021) 4.

[HW12] HOSEK, LUKAS and WILKIE, ALEXANDER. "An analytic model for full spectral sky-dome radiance". *ACM Transactions on Graphics (TOG)* 31.4 (2012), 1–9 2.

[JS20] JEFFREY OKUN, VES and SUSAN ZWERMAN, VES. *The VES handbook of visual effects: industry standard VFX practices and procedures*. Routledge, 2020 2.

[JSRV22] JAKOB, WENZEL, SPEIERER, SÉBASTIEN, ROUSSEL, NICOLAS, and VICINI, DELIO. "Dr. jit: A just-in-time compiler for differentiable rendering". *ACM Transactions on Graphics (TOG)* 41.4 (2022) 7.

[KHFH11] KARSCH, KEVIN, HEDAU, VARSHA, FORSYTH, DAVID, and HOIEM, DEREK. "Rendering synthetic objects into legacy photographs". *ACM Transactions on graphics (TOG)* 30.6 (2011), 1–12 2.

[KRFB06] KHAN, ERUM ARIF, REINHARD, ERIK, FLEMING, ROLAND W, and BÜLTHOFF, HEINRICH H. "Image-based material editing". *ACM Transactions on Graphics (TOG)* 25.3 (2006), 654–663 6, 7, 9.

[LHL*24] LIN, ZHI-HAO, HUANG, JIA-BIN, LI, ZHENGQIN, et al. "IRIS: Inverse Rendering of Indoor Scenes from Low Dynamic Range Images". *arXiv preprint arXiv:2401.12977* (2024) 2.

[LLYL20] LIU, YUNFEI, LI, YU, YOU, SHAODI, and LU, FENG. "Unsupervised learning for intrinsic image decomposition from a single image". *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, 3248–3257 2, 3, 7.

[LMF*19] LEGENDRE, CHLOE, MA, WAN-CHUN, FYFFE, GRAHAM, et al. "Deeplight: Learning illumination for unconstrained mobile mixed reality". *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, 5918–5928 2.

[LYO*23] LI, ZHENGQIN, YU, LI, OKUNEV, MIKHAIL, et al. "Spatiotemporally consistent hdr indoor lighting estimation". *ACM Transactions on Graphics* 42.3 (2023), 1–15 3.

[MBHD18] MARNERIDES, DEMETRIS, BASHFORD-ROGERS, THOMAS, HATCHETT, JONATHAN, and DEBATTISTA, KURT. "Expandnet: A deep convolutional neural network for high dynamic range expansion from low dynamic range content". *Computer Graphics Forum*. Vol. 37. 2. 2018, 37–49 3.

[MKC*17] MAIER, ROBERT, KIM, KIHWAN, CREMERS, DANIEL, et al. "Intrinsic3D: High-quality 3D reconstruction by joint appearance and geometry optimization with spatially-varying lighting". *Proceedings of the IEEE international conference on computer vision*. 2017, 3114–3122 2.

[MMB12] MITTAL, ANISH, MOORTHY, ANUSH KRISHNA, and BOVIK, ALAN CONRAD. "No-Reference Image Quality Assessment in the Spatial Domain". *IEEE Transactions on Image Processing* 21.12 (2012), 4695–4708 7.

[PCS*24] PHONGTHAWEE, PAKKAPON, CHINCHUTHAKUN, WORAMETH, SINSUNTHITHET, NONTAPHAT, et al. "Diffusionlight: Light probes for free by painting a chrome ball". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, 98–108 2, 3.

[PMGD21] PHILIP, JULIEN, MORGENTHALER, SÉBASTIEN, GHARBI, MICHAËL, and DRETTAKIS, GEORGE. "Free-viewpoint indoor neural relighting from multi-view stereo". *ACM Transactions on Graphics (TOG)* 40.5 (2021), 1–18 2.

[RBL*22] ROMBACH, ROBIN, BLATTMANN, ANDREAS, LORENZ, DOMINIK, et al. "High-resolution image synthesis with latent diffusion models". *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, 10684–10695 3, 4.

[SJT*24] SHAH, UZAIR, JASHARI, SARA, TUKUR, MUHAMMAD, et al. "VISPI: Virtual Staging Pipeline for Single Indoor Panoramic Images". *STAG 2024*. Eurographics Association. 2024 3.

[SMT*20] SRINIVASAN, PRATUL P, MILDENHALL, BEN, TANCIK, MATTHEW, et al. "Lighthouse: Predicting lighting volumes for spatially-coherent illumination". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, 8080–8089 2, 3.

[SYMH23] SHIN, GYEONGIK, YU, KYEONGMIN, MARK, MPABULUNGI, and HONG, HYUNKI. "Hdr map reconstruction from a single ldr sky panoramic image for outdoor illumination estimation". *IEEE Access* 11 (2023), 17359–17374 3.

[WCA*22] WANG, ZIAN, CHEN, WENZHENG, ACUNA, DAVID, et al. "Neural light field estimation for street scenes with differentiable virtual object insertion". *European Conference on Computer Vision*. Springer. 2022, 380–397 3, 7, 8.

[WCL*23] WANG, JIONGHAO, CHEN, ZIYU, LING, JUN, et al. "360-degree panorama generation from few unregistered nfov images". *arXiv preprint arXiv:2308.14686* (2023) 2–4.

[WGL22] WEBER, HENRIQUE, GARON, MATHIEU, and LALONDE, JEAN-FRANÇOIS. "Editable indoor lighting estimation". *European Conference on Computer Vision*. Springer. 2022, 677–692 2, 3, 6, 7, 9.

[WPFK21] WANG, ZIAN, PHILION, JONAH, FIDLER, SANJA, and KAUTZ, JAN. "Learning indoor inverse rendering with 3d spatially-varying lighting". *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, 12538–12547 3.

[WSP*23] WANG, CHAO, SERRANO, ANA, PAN, XINGANG, et al. "Glowgan: Unsupervised learning of hdr images from ldr images in the wild". *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, 10509–10519 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, 3, 6, 7, 9.

[YAH*23] YU, HONG-XING, AGARWALA, SAMIR, HERRMANN, CHARLES, et al. "Accidental light probes". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, 12521–12530 2.

[YKH*24] YANG, LIHE, KANG, BINGYI, HUANG, ZILONG, et al. "Depth anything: Unleashing the power of large-scale unlabeled data". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, 10371–10381 2–5, 7.

[YLL*21] YU, HANNING, LIU, WENTAO, LONG, CHENGJIANG, et al. "Luminance attentive networks for HDR image and panorama reconstruction". *Computer Graphics Forum*. Vol. 40. 7. 2021, 181–192 2–4.

[ZCC16] ZHANG, EDWARD, COHEN, MICHAEL F., and CURLESS, BRIAN. "Emptying, refurnishing, and relighting indoor spaces". *ACM Transactions on Graphics*. 2016, 1–14 2.

[ZL17] ZHANG, JINSONG and LALONDE, JEAN-FRANÇOIS. "Learning high dynamic range from outdoor panoramas". *Proceedings of the IEEE International Conference on Computer Vision*. 2017, 4519–4528 3.

[ZRA23] ZHANG, LVMIN, RAO, ANYI, and AGRAWALA, MANEESH. "Adding conditional control to text-to-image diffusion models". *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, 3836–3847 3, 4.

[ZSH*19] ZHANG, JINSONG, SUNKAVALLI, KALYAN, HOLD-GEOFFROY, YANNICK, et al. "All-Weather Deep Outdoor Lighting Estimation". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2019 2, 6, 7, 9.

[ZYZ*22] ZHAN, FANGNENG, YU, YINGCHEN, ZHANG, CHANGGONG, et al. "Gmlight: Lighting estimation via geometric distribution approximation". *IEEE Transactions on Image Processing* 31 (2022) 2.

[ZZY*21] ZHAN, FANGNENG, ZHANG, CHANGGONG, YU, YINGCHEN, et al. "Emlight: Lighting estimation via spherical distribution approximation". *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 4. 2021, 3287–3295 2.