Hybrelighter: Combining Deep Anisotropic Diffusion and Scene Reconstruction for On-device Real-time Relighting in Mixed Reality

Hanwen Zhao, John Akers, Baback Elmieh, Ira Kemelmacher-Shlizerman

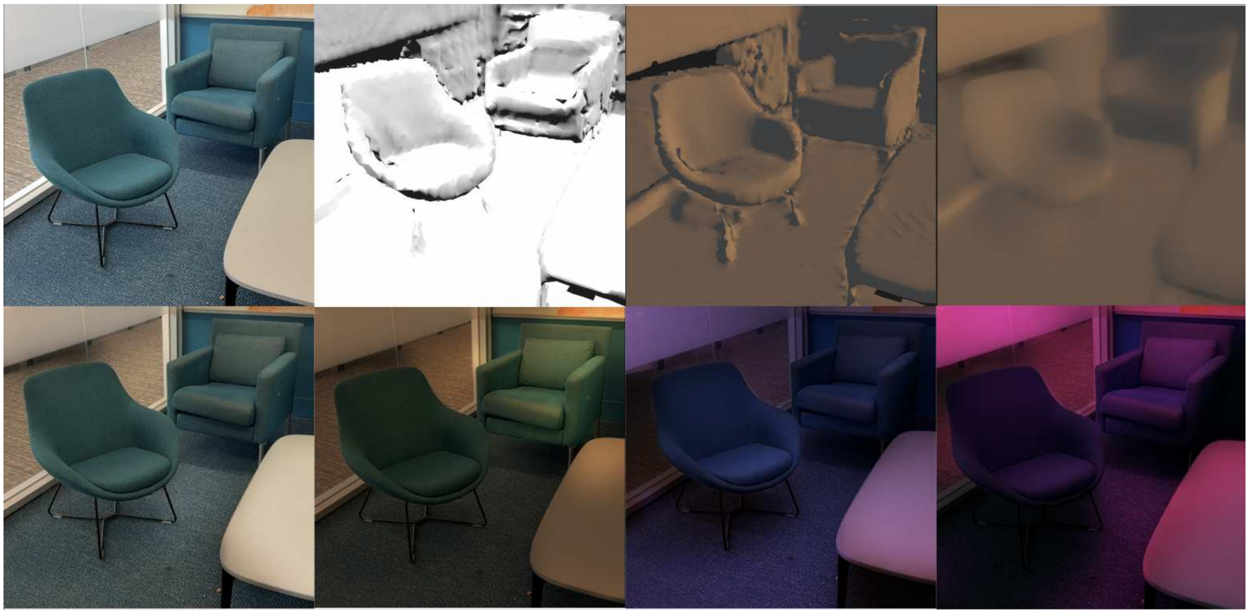


Fig. 1: Utilize the scene reconstruction capability and anisotropic diffusion (row 1) for real-time relighting of a room in various lighting conditions in mixed reality (row 2).

Abstract—Mixed Reality scene relighting, where virtual changes to lighting conditions realistically interact with physical objects, producing authentic illumination and shadows, can be used in a variety of applications. One such application in real estate could be visualizing a room at different times of day and placing virtual light fixtures. Existing deep learning-based relighting techniques typically exceed the real-time performance capabilities of current MR devices. On the other hand, scene understanding methods, such as on-device scene reconstruction, often yield inaccurate results due to scanning limitations, in turn affecting relighting quality. Finally, simpler 2D image filter-based approaches cannot represent complex geometry and shadows. We introduce a novel method to integrate image segmentation, with lighting propagation via anisotropic diffusion on top of basic scene understanding, and the computational simplicity of filter-based techniques. Our approach corrects on-device scanning inaccuracies, delivering visually appealing and accurate relighting effects in real-time on edge devices, achieving speeds as high as 100 fps. We show a direct comparison between our method and the industry standard, and present a practical demonstration of our method in the aforementioned real estate example.

Index Terms—Relighting, Mixed Reality, Computer Vision

1 Introduction

Recent advancements in Mixed Reality (MR) technologies have significantly enhanced the capabilities of headsets, enabling them not only to visualize but also to comprehend real-world environments through passthrough cameras. This capability allows users to interact with their environments through a controllable virtual rendering layer, opening opportunities for various applications, notably scene relighting. By intelligently manipulating camera data, we can effectively alter the perceived lighting conditions of the environment, offering realistic simulations of diverse lighting scenarios without physically altering actual light sources.

Scene relighting has numerous practical applications, including immersive integration of virtual objects with real-world scenes, creating visually compelling storytelling effects, or simulating environments under different lighting conditions for planning and visualization.

Simple 2D filter-based relighting methods, such as adjustments to image temperature, contrast, or exposure, can be easily implemented with various image editing software packages. More advanced techniques involve masking-based image processing, which allows localized illumination changes, like simulating a spotlight effect. These learning-free approaches are highly efficient, stable, and predictable but suffer from limitations due to the absence of 3D geometric context, particularly when accurately representing complex shadows and occlusion interactions.

In contrast, sophisticated 3D-based relighting methods utilize scene geometry derived from real-time reconstruction to achieve more accurate lighting simulations. These techniques leverage traditional ren-

- *Hanwen Zhao, University of Washington, Seattle, Washington, United States, hzhao5@uw.edu*
John Akers, UW Reality Lab, University of Washington, Seattle, Washington, United States
Baback Elmieh, Computer Science, University of Washington, Seattle, Washington, United States, baback@cs.washington.edu
Prof Ira Kemelmacher-Shlizerman, University of Washington , Seattle , Washington, United States

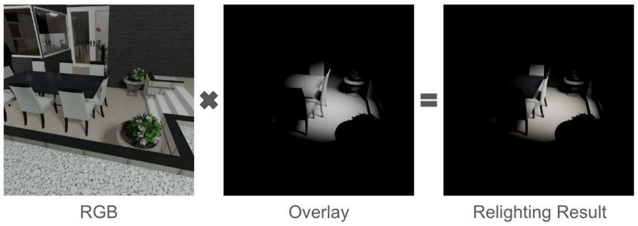


Fig. 2: An example of a synthetic scene being relit by a spotlight image filter rendered on an untextured version of the mesh data.

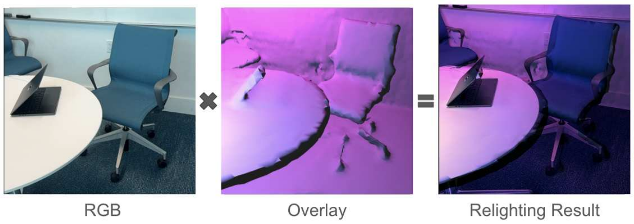


Fig. 3: An example of the mesh quality directly from the LiDAR camera on the iPhone 16 Pro. The errors in such suboptimal mesh can significantly reduce the visual quality of the final relit image, especially around the edges.

dering methods, such as rasterization or ray tracing, enabling more realistic representation of shadows and illumination interactions. However, current on-device scene reconstruction, typically performed using built-in lidar or depth sensors on MR headsets, often results in simplified meshes due to computational constraints and limited fidelity, adversely affecting the quality of relighting.

More recent developments in deep learning have paved the way for highly detailed and realistic relighting solutions. Deep neural networks can infer intrinsic scene properties such as surface normals, roughness, albedo, and reflectance, greatly improving the accuracy and visual realism of relighting effects. However, these deep learning methods are typically computationally intensive, limiting their real-time applicability on edge devices such as MR headsets.

In this paper, we propose a novel hybrid method that integrates the computational efficiency of filter-based approaches with the deeper semantic understanding provided by deep learning-guided anisotropic diffusion. Our approach specifically addresses the inaccuracies present in simplified meshes generated by real-time scene reconstruction. By leveraging RGB images captured from passthrough cameras, we employ a deep learning-based feature extractor to accurately identify high-frequency details and object boundaries. Subsequently, we apply anisotropic diffusion iteratively, guided by these learned features, to smoothly interpolate shading within object boundaries while maintaining sharp, precise edges.

Recognizing the computational demands of conventional anisotropic diffusion, which typically require numerous iterations to achieve equilibrium, we introduce a cascaded diffusion strategy. This method capitalizes on the rapid propagation of gradients at lower image resolutions, progressively refining edge details through fewer iterations, significantly enhancing runtime efficiency.

In short, the problem we are solving is: How can we relight a dynamically updating scene that is geometrically correct, and runs in real-time on device in Mixed Reality? And our primary contributions include:

- A hybrid relighting approach that effectively combines deep learning feature extraction with filter-based anisotropic diffusion.
- A modified anisotropic diffusion process optimized for fewer iterations, demonstrating that existing deep diffusion model architectures can be effectively adapted to run in real time on edge devices.
- A practical demonstration and evaluation of our method on commercially available MR devices. We present how relighting can be practically applied in a real estate tour scenario.

2 Related Work

Scene Reconstruction based Relighting Approaches

Scene relighting techniques have historically relied on high-quality reconstructions as a basis for realistic lighting simulations. Early research demonstrated the potential of mesh-based methods to produce convincing relit images given sufficiently detailed geometry [9]. However, due to computational constraints, MR headsets typically employ mesh decimation strategies, resulting in reduced fidelity and diminished relighting quality. To address this limitation, several studies have proposed methods to incrementally refine meshes dynamically and efficiently. Hybrid voxel-octree fusion techniques [14], and the Large Reconstruction Model (LRM) [22], demonstrate that high-quality meshes can be rapidly generated even from sparse input views. Additionally, recent advancements in real-time dense scene reconstruction can achieve high-fidelity meshes at approximately 20 frames per second (fps) on powerful GPUs, although these methods remain computationally prohibitive for edge-device deployment [15].

Neural rendering methods, including Gaussian splatting and Neural Radiance Fields (NeRF), have also been explored for their impressive capabilities in large-scale scene relighting [16] [3] [26]. Despite achieving visually compelling results, these methods require extensive pre-training, ranging from several minutes to half an hour, thus limiting their practical usability in interactive MR scenarios.

Other recent approaches seek to infer 3D scene information without explicitly generating 3D representations. Advances in depth estimation have enabled consistent depth image generation from extended video sequences, achieving high-performance processing speeds of approximately 9 ms per frame (over 100 fps) on high-end discrete GPUs [4]. Additionally, depth estimation methods tailored for edge devices have been developed, reaching speeds up to 300 fps on embedded Nvidia hardware, albeit with compromised image quality and temporal inconsistency [6].

An approximation of surface normals can be efficiently derived using camera intrinsics and pixel-wise cross-products; however, accurately capturing complex occlusion information necessary for realistic shadow rendering introduces additional computational overhead, oftentimes requiring another learning-based model running on top of it as explored in the work by Yang et al. [23]. The quality of relighting thus significantly depends not only on global depth consistency across frames but also on local depth accuracy. To address this, joint prediction frameworks for depth and surface normals have emerged, demonstrating superior quality but sacrificing real-time performance on edge computing devices [8].

Image-based Direct Relighting Methods

Image-based relighting has been extensively investigated within computer vision literature. Traditional techniques typically involve decomposing images into intrinsic components, such as albedo, shading, and lighting, often requiring multiple neural network layers for accurate extraction [24] [11]. More recent research leverages generative models capable of directly synthesizing relit images conditioned upon new lighting parameters, achieving high-quality results with simpler architectures. These generative models demonstrate remarkable flexibility and visual realism [18] [2] [10]. Although explicit runtime evaluations are not always provided, related studies on diffusion-based generative methods indicate that one-step diffusion models can produce relit images as rapidly as 9 ms per frame on high-end GPUs [25]. However, all these methods are tailored for static images, without considering temporal information which makes maintaining consistent illumination across video sequences in real-time a challenge requiring further investigation.

Guided Depth and Point Cloud Enhancement

Working with low-fidelity data from depth sensors and lidar scanners is a significant challenge in many fields, including MR, autonomous

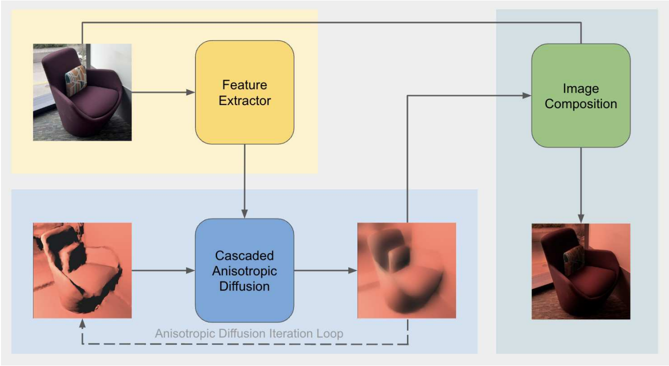

Feature Extractor
Image Composition
Cascaded Anisotropic Diffusion
Anisotropic Diffusion Iteration Loop

Fig. 4: Our relighting pipeline takes in one RGB camera image and one RGB relight image. High-level edge information extracted from the RGB camera image is used as the guidance for anisotropic diffusion running on the relight image. The refined output is composited with the original camera RGB to produce the final relit image.

driving, and robotics. Thus having high-quality depth and point cloud data are crucial for accurate scene understanding and effective relighting. Recent approaches have leveraged concurrent RGB imagery to enhance [17] [5] [28] [20] and complete sparse or imperfect depth information [19] [21], exploiting complementary information between modalities. Similar approaches can also be applied on point-cloud completion [13] [12]. Studies on guided depth super-resolution and depth completion indicate that real-time performance (up to 50 fps) is achievable, highlighting the potential of these methods for practical MR deployment [7].

Our proposed method draws inspiration from this body of work. Specifically, we combine RGB guidance and suboptimal mesh data to significantly improve scene relighting quality. Our approach employs anisotropic diffusion guided by learned RGB features, effectively translating established depth super-resolution methods into the context of real-time MR relighting.

3 Methods

Our proposed relighting method integrates a mesh-aware, filter-based approach with guided anisotropic diffusion, delivering high-quality, real-time relighting for mixed reality (MR) scenarios on edge devices. This method combines the speed and efficiency of GPU-based filtering with the semantic precision of deep learning-based anisotropic diffusion, effectively overcoming the visual quality limitations introduced by simplified mesh reconstruction typically found in MR hardware. The following sections detail each component of our pipeline, including mesh-aware filtering, guided anisotropic diffusion, cascaded diffusion strategies for enhanced computational efficiency, shadow rendering adjustments, and transferable training methodologies derived from existing depth super-resolution techniques.

3.1 Mesh-aware Filter-based Relighting

To leverage the computational speed of simple 2D-filters we utilize a mesh-aware filtering pipeline that integrates geometric awareness into traditional image processing. Initially, we reconstruct the scene geometry using real-time mesh reconstruction techniques typically available on MR devices. Within a 3D rendering engine, we introduce virtual lights and render the mesh using standard shading algorithms. The resulting shaded mesh acts as an image-space filter that is precisely aligned with the RGB camera frames through consistent camera intrinsic parameters. The final relit image is obtained by compositing this rendered mesh with the original RGB frame via image multiplication.

Thus, it can be defined as the following: we are given a rendered relight image $R \in \mathbb{R}^{H\times W\times 3}$ serving as the filter acquired directly from the original mesh. And we are given an RGB camera frame $C \in \mathbb{R}^{H\times W\times 3}$ as the passthrough. The output which is our relit image which can be simply defined as

$$S = R \odot C \tag{1}$$

where $\odot$ denotes element-wise (per-pixel) multiplication. While this approach provides efficiency and real-time performance, it depends significantly on the accuracy and fidelity of the reconstructed mesh. Due to the computational limitations inherent in MR hardware, meshes are often simplified, resulting in degraded visual quality and inaccurate shading outcomes.

3.2 Guided Anisotropic Diffusion for Relighting Correction

Our key insight is that the mesh inaccuracies that affect shading quality most severely occur at 3D depth discontinuities which correspond to edge boundaries in 2D images. To solve for this, we employ anisotropic diffusion, an edge-aware filtering technique known for effectively smoothing interior regions of objects while preserving edges. Specifically, we guide anisotropic diffusion using edge information extracted by a deep learning-based feature extractor trained to identify critical scene details and object boundaries. By using learned features rather than simple RGB edges, we enhance diffusion precision, ensuring the gradient propagation remains confined within object boundaries.

Drawing inspiration from guided depth super-resolution research [17], we adapt the anisotropic diffusion approach from the depth domain to the relighting scenario. The guided anisotropic diffusion can be mathematically described using the formulation introduced by Metzger et al. [17]. The prediction of a pixel's value at iteration t at location p is expressed as follows:

$$\hat{y}_t^p = y_{t-1}^p + \lambda \cdot \sum_{n\in\mathcal{N}_4^p} \left(y_{t-1}^n - y_{t-1}^p\right) \cdot c(\mathbf{g}^p, \mathbf{g}^n) \tag{2}$$

Here, y_t^p represents the pixel intensity at position p in the image $\mathbf{Y}_t$. $\mathcal{N}_4^p$ refers to the set of four directly adjacent pixels (4-neighborhood) around pixel p, effectively forming a planar graph across the image. The hyperparameter λ controls the diffusion strength and ensures stability during iterations; when utilizing 4-neighborhood connectivity, it must satisfy $\lambda < \frac{1}{4}$. The diffusion coefficient $c(\mathbf{g}^p, \mathbf{g}^n)$ is computed based on the similarity between guide features at pixels p and n, as initially introduced by [17]:

$$c(\mathbf{g}^p, \mathbf{g}^n) = \frac{\kappa^2}{\kappa^2 + \|\mathbf{g}^p - \mathbf{g}^n\|_2^2} \tag{3}$$

The hyperparameter κ controls sensitivity to feature gradients within the guidance image $\mathbf{G}$, with higher values allowing smoother diffusion across larger differences. The symmetry of this coefficient function ensures $c(\mathbf{g}^p, \mathbf{g}^n) = c(\mathbf{g}^n, \mathbf{g}^p)$.

While conventional anisotropic diffusion methods directly use the current image state $\mathbf{Y}_{t-1}$ as guidance, the method we adopt explicitly incorporates separate guidance data $\mathbf{G}$. This distinction is particularly valuable in relighting tasks. Unlike depth maps, relit images often contain inaccuracies due to simplifications in the initial mesh representation. Thus, our method addresses these inherent inaccuracies explicitly, optimizing the diffusion process to yield visually consistent results in fewer iterations.

3.3 Cascaded Anisotropic Diffusion for Enhanced Efficiency

Recognizing the computational burden of conventional anisotropic diffusion, we introduce a cascaded diffusion strategy operating at multiple

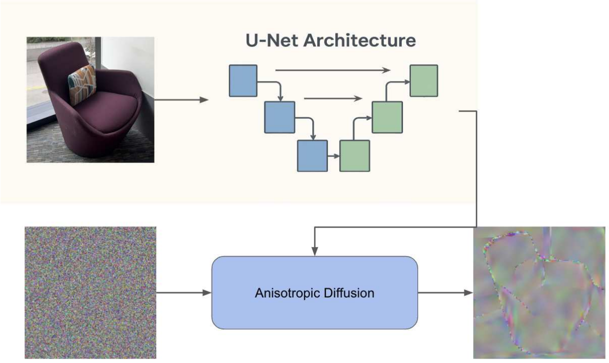


Fig. 5: The effect of the anisotropic diffusion coefficients provided by the RGB camera guidance image. The input for anisotropic diffusion is pure Gaussian noise. The output is scaled up for better image clarity.

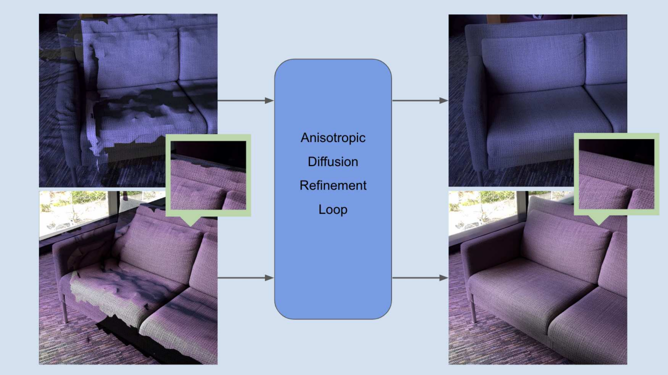


Fig. 6: Edge-aware pixel-color diffusion process demonstrates strong capabilities in fixing the errors caused by inaccuracies in the scanning process. We show how this approach can produce smooth shading in each object while maintaining high edge consistency.

resolutions.

(1) Initialize coarse level:

$$\mathbf{Y}_0^{(1)} = \downarrow^{L-1}(\mathbf{Y}_0), \quad \mathbf{G}^{(1)} = \downarrow^{L-1}(\mathbf{G})$$

(2) Coarse-to-fine diffusion:

For $l = 1$ to L :

$$\mathbf{C}^{(l)} = \text{MinPool}\left(c(\mathbf{G}^{(l)})\right)$$

$$\mathbf{Y}_t^{(l)} = \mathscr{D}\left(\mathbf{Y}_{t-1}^{(l)}, \mathbf{C}^{(l)}\right)$$

(3) Upsample and refine:

$$\text{If } l < L: \quad \mathbf{Y}_t^{(l+1)} = \uparrow\left(\mathbf{Y}_t^{(l)}\right)$$

(4) Final relit image:

$$\mathbf{S} = \mathbf{Y}_t^{(L)} \tag{4}$$

In this formulation, let L be the number of resolution levels, where $l = 1$ denotes the coarsest scale and $l = L$ corresponds to the original resolution. The image $\mathbf{Y}_t^{(l)}$ and guide $\mathbf{G}^{(l)}$ are derived by downsampling the original inputs $\mathbf{Y}_0$ and $\mathbf{G}$ respectively. At each level, we compute diffusion coefficients $\mathbf{C}^{(l)}$ using a min pooling strategy, ensuring edge preservation across pooling grids. The anisotropic diffusion operator $\mathscr{D}$ is then applied using these coefficients. The diffused image is upsampled to the next finer level, where additional refinement iterations are performed. This coarse-to-fine cascade continues until the finest resolution is reached, yielding the final relit image $\mathbf{S}$. This approach significantly reduces computational overhead while preserving critical high-frequency details needed for sharp visual relighting. By initially performing diffusion at reduced image resolutions, we significantly expedite gradient propagation across large pixel areas. We employ min pooling for downsampling diffusion coefficients, ensuring edge integrity by treating any edge pixel as an edge across the entire pooling grid, thereby preventing unintended color blending.

Subsequently, we progressively upscale the image back to its original resolution, applying additional, targeted diffusion iterations at each resolution level to restore detailed edge information lost during downsampling. This cascading approach significantly reduces computational requirements, enabling real-time performance while preserving critical high-frequency details necessary for visually sharp relighting.

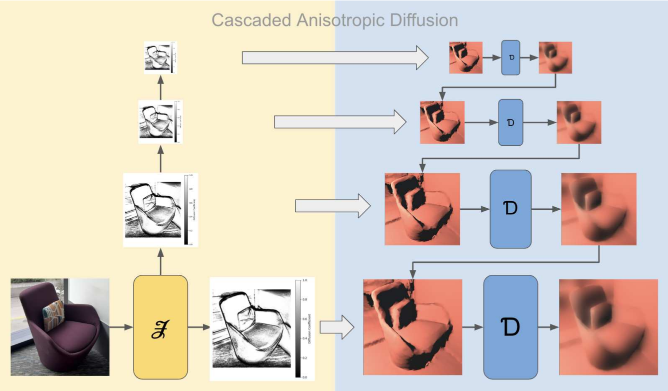


Fig. 7: Our cascaded anisotropic diffusion pipeline. We run the diffusion process at lower resolution and we chain the results together by upsampling the previous outputs and further refining image quality.

3.4 Adjusted Anisotropic Diffusion for Shadow Rendering

An important aspect of realistic relighting is accurately rendering shadows cast by real objects. Unlike direct shading, shadows' shapes depend on the geometry of casting objects rather than the surfaces onto which they fall. To preserve the accuracy and softness of shadow edges, we implement a separate diffusion pass specifically optimized for shadow rendering, as shown in Fig. 8. By reducing the number of diffusion iterations and maintaining operation at higher resolutions, we effectively blur shadows to produce soft, realistic effects without compromising shape accuracy. The edge-aware nature of the anisotropic diffusion process ensures that shadow colors remain confined, preventing leakage onto adjacent objects.

3.5 Transferable Training from Depth Super-resolution Models

The core learnable component of our method is the deep learning-based feature extractor responsible for guiding anisotropic diffusion. Leveraging existing depth super-resolution training frameworks, we train a compact, state-of-the-art model suitable for edge-device deployment. Our training pipeline modifies the input channels from single-channel depth maps to three-channel RGB images, facilitating the direct application of existing depth datasets by artificially randomizing color tints. This approach significantly simplifies the training process, benefiting from widely available depth datasets while addressing the scarcity of annotated relighting datasets.

4 Experiments

4.1 Training Details

The trainable component within our relighting pipeline is the deep feature extractor used on RGB camera images. An effective feature extractor ideally produces values close to zero at edges and close to

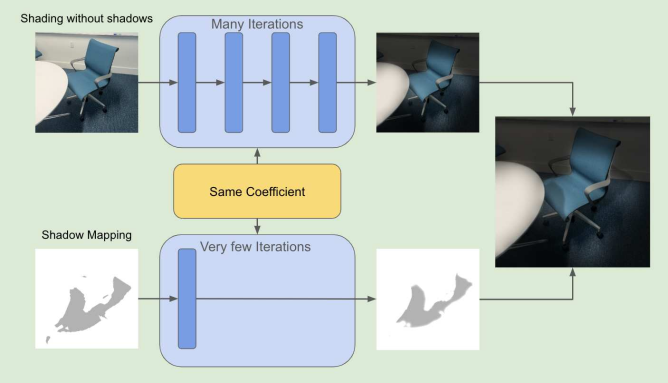


Fig. 8: Shadows can be computed in a separate pass using traditional shadow mapping techniques together with only a few of anisotropic diffusion iterations to ensure their overall shapes are not destructed by the diffusion process.

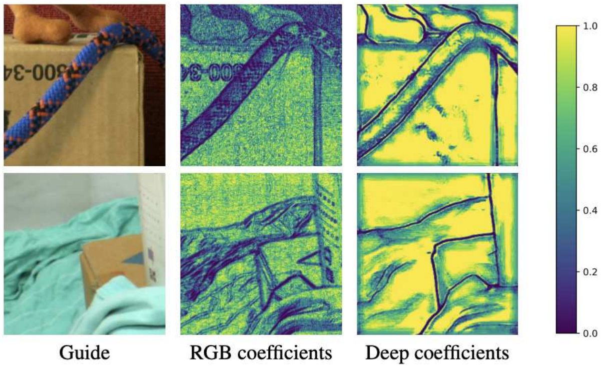


Fig. 9: A figure from [17]. This figure directly visualizes the values of the coefficients. We observe that they have the exact properties we want for guided anisotropic diffusion. Since this is the only trainable section in the whole pipeline, this shows that any well trained deep feature extractor can be used for any guided anisotropic diffusion tasks.

one within object regions, facilitating smooth gradient propagation internally while preserving sharp boundaries. We observe this characteristic to be universally beneficial regardless of the specific downstream tasks, as the optimization consistently targets the diffusion coefficients. Consequently, our training procedure is significantly inspired by the approach detailed in Metzger et al. [17].

We use widely accessible datasets from the domain of depth super-resolution for training, specifically the DIML dataset. Our experiments involve training two state-of-the-art mobile-oriented vision models: mobilenetv4_conv_small and mobilenetv4_conv_medium. Both models employ a U-Net structure from the segmentation_models_pytorch package to generate precise feature maps.

A notable departure from the original method presented by Metzger et al. is that our feature extraction exclusively relies on RGB images, omitting the concatenation of lower-resolution depth inputs. This change is essential because the relighting images available to us inherently contain inaccuracies and inconsistencies, thus precluding their use as reliable ground truth even at reduced resolutions. Removing these erroneous inputs ensures that our models do not learn to depend on potentially faulty data. Another significant innovation introduced by Metzger et al., known as "adjustment step," is crucial for ensuring convergence during training on depth datasets. We also integrate this technique into our training procedure to improve stability and accuracy.

Lastly, we clearly outline our training parameters, highlighting key adjustments from the original configuration. We adopt the same dynamically stepped learning rate as proposed by Metzger et al. but use higher-resolution 512x512 images to ensure enhanced edge fidelity. The depth upsampling scale factor is set at 32. Due to memory constraints during backpropagation and data tracking, we configure the training step as 512 and predictions as 20,000. Both vision models are trained on NVIDIA A100 GPUs, with a batch size of 8 over 12,000 iterations.

4.2 Numerical Metrics

We conduct comprehensive evaluations of our method using the ARKitScenes dataset [1], which is collected by Apple using the same LiDAR camera embedded within their mixed reality devices. This dataset comprises RGB camera frames, ARKit-generated meshes identical to those accessible on-device, and high-density point clouds acquired by a Faro laser scanner, providing an ideal benchmark for assessing our relighting approach.

Given that the detailed point cloud data surpasses the real-time processing capabilities of edge devices, we perform offline rendering of both the ARKit-generated meshes and the high-density point clouds using Blender. This ensures alignment consistency between the rendered meshes and the RGB camera frames. Subsequently, we employ a mesh-based filtering strategy in Blender by adding virtual light sources directly onto both meshes and point clouds. The rendered outputs from this step are composited with the corresponding RGB camera frames to generate final relit images, ensuring accurate alignment by utilizing the provided pose and trajectory information.

To assess our method, we introduce three benchmarks, each designed to evaluate distinct aspects of our pipeline:

Benchmark 1: Mesh Error Correction

We first evaluate our method's effectiveness at correcting inaccuracies inherent in the ARKit mesh. In this scenario, we simulate an ideal lighting condition by placing a bright point light centrally within the room. A flawless mesh rendering would yield a grey to white filter, implying low errors in the composited result. Conversely, an imperfect mesh introduces black pixels in the filter, resulting in artifacts in the final composited image. We quantitatively measure performance using both LPIPS [27] and Peak Signal to Noise Ratio (PSNR) metrics, comparing:

- Original RGB frames and directly composited ARKit mesh-based rendering
- Original RGB frames and our refined anisotropic diffusion-enhanced rendering

Our results demonstrate that our refined method achieves lower LPIPS and higher PSNR scores, verifying its capability to effectively mitigate artifacts caused by mesh inaccuracies.

Benchmark 2: Multi-Lighting Consistency

Acknowledging that trivial solutions (such as a uniformly white filter) could artificially perform well in Benchmark 1, we introduce a second scenario that mimics realistic relighting conditions more closely. Here, two dimmed, differently-colored point lights are placed within the scene to generate a different lighting environment. Again, we assess the rendered images against original RGB frames using LPIPS and PSNR metrics. Our refined method consistently exhibits lower LPIPS scores compared to direct mesh rendering, emphasizing its ability to retain the visual features of the original scenes under varied lighting configurations.

Benchmark 3: Relighting Fidelity

Lastly, we evaluate our method's capability to generate visually accurate relighting effects relative to a high-fidelity baseline rendered from dense point cloud data. By comparing:

- Our refined method vs. the high-fidelity point cloud rendering
- Direct mesh rendering vs. the high-fidelity point cloud rendering

Although anisotropic diffusion can introduce slight pixel color differences due to inherent smoothing, as reflected in the PSNR scores, our approach demonstrates strong performance in preserving structural consistency and overall visual realism, as evidenced by lower LPIPS scores. This outcome aligns with our intuition: since the anisotropic

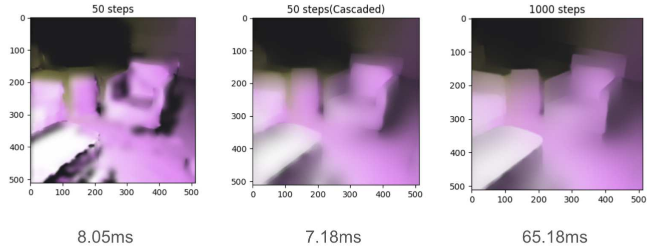


Fig. 10: Cascaded diffusion can quickly propagate gradients for high visual quality at a fraction of the time to run.

diffusion result serves primarily as a filter that preserves edges while smoothing other areas, compositing it with the original, sharp RGB image mitigates potential noise introduced when smoothing was performed. Therefore, the clarity and sharpness of the original RGB image ensure that the final composited output maintains high visual fidelity.

For the metrics section, all data was gathered using the mobilenetv4_conv_medium model, with cascaded anisotropic diffusion running with 10 steps at 32×32 resolution, 15 steps at 64×64, 25 steps at 128×128, 30 steps at 256×256, and then up-sampled to 512×512 resolutions. Quantitative results from these benchmarks are summarized in Tab. 1, accompanied by representative visual comparisons illustrated in Fig. 14, Fig. 15, Fig. 16. These evaluations collectively underscore our approach's robustness, accuracy, and practical applicability for real-time MR relighting tasks.

Table 1: Evaluation metrics for relighting methods on mesh ID 47333462 with 596 images in the ARKitScenes dataset. Lower LPIPS and higher PSNR values indicate better performance. Bold indicates the better (smaller) value for each metric in each benchmark.

Benchmark	ARKitMesh	Ours
Mesh Error Correction		
-LPIPS	0.1322	**0.1078** 18% ↓
-PSNR	12.7437	**12.8100** 0.5% ↑
Multi-Lighting Consistency		
-LPIPS	0.4310	**0.4120** 4% ↓
-PSNR	**1.7916**	1.7672 1.4% ↓
Relighting Fidelity		
-LPIPS	0.1391	**0.1200** 14% ↓
-PSNR	**15.7850**	15.7437 0.3% ↓

4.3 Ablations

In this section, we build upon the analyses presented by Metzger et al [17]. in their training details by specifically investigating key components unique to our inference-time techniques.

Effectiveness of the Cascaded Diffusion Approach

The cascaded diffusion strategy is engineered primarily to reduce the number of diffusion iterations, thereby achieving real-time performance suitable for deployment on edge devices. Here, we rigorously assess its effectiveness and quantify the speed-up it offers. We employ the medium-sized MobileNet model, applying a cascaded diffusion sequence comprising 5 steps at 32×32 resolution, 10 steps at 64×64, 15 steps at 128×128, 20 steps at 256×256, and then up-sampled to 512×512 resolutions. This configuration totals 50 diffusion steps distributed across multiple scales. We compare this cascaded strategy against two baseline scenarios: 50 and 1000 diffusion steps, all executed solely at the resolution of 256×256 and up-sampled to 512x512. We also measure the runtime performance for each scenarios. The result is shown in Fig. 10.

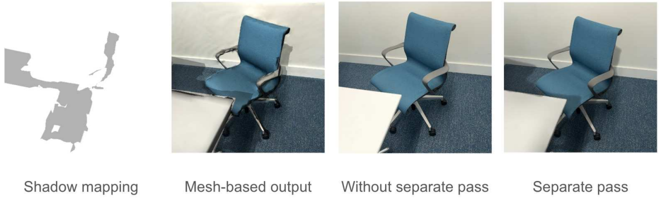


Fig. 11: A naive diffusion approach can fail when the shadow is cast onto simple flat surfaces such as walls or floors. A separated pass for shadow rendering can effectively retain the overall shape.

Separate Rendering Pass for Shadows

We further examine the impact of conducting the diffusion process directly on relighting images that include pre-applied shadows, compared to executing the diffusion on shadow-free relighting images and subsequently applying shadows in a separate rendering pass. In Fig. 11, we illustrate how directly diffusing images with embedded shadows significantly deteriorates shadow fidelity, primarily due to lack of strict geometric confinement. Conversely, applying shadows separately and refining through targeted shadow map iterations effectively preserves and enhances shadow quality, mitigating the degradation introduced by direct diffusion.

4.4 Use Case Demonstration

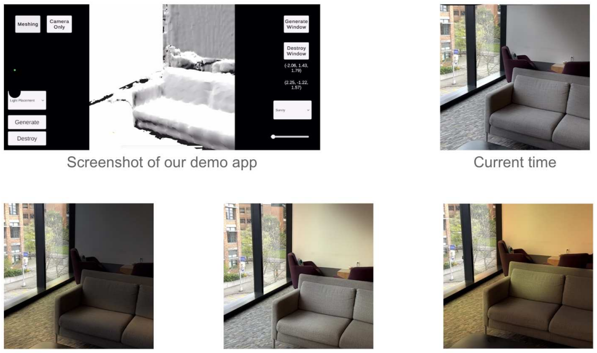
To demonstrate our approach in a practical mixed reality context, we created a Unity-based demo targeted primarily at handheld devices, such as iPads, suitable for real-estate tour scenarios. Our demo leverages ARKit's scene reconstruction capabilities to swiftly scan the environment, allowing users to define the positions and orientations of windows through intuitive rectangular area markers. Subsequently, users can interactively visualize different lighting scenarios across various times of the day using a slider interface. This capability remains effective regardless of the actual ambient lighting conditions at runtime, provided the environment is reasonably illuminated. Additionally, the demo includes scenarios such as cloudy day lighting, offering flat, uniform illumination, and nighttime illumination featuring moonlight effects.

An important extension of our demo addresses custom virtual lighting configurations, highly relevant in practical scenarios such as virtual product demonstrations or personalized interior design planning. Users can intuitively place various virtual light sources, including lamps, candles, or decorative lighting, directly within the real-world environment. The system then dynamically calculates and visualizes realistic lighting interactions and shadows, enabling users to evaluate aesthetic and practical implications without physically altering the actual lighting setups.

Moreover, our demo supports fully dynamic mesh updating, which significantly enhances user interactivity and realism. Users can freely rearrange furniture or other elements within the room and immediately observe the updated lighting effects. This dynamic capability allows real-time assessment of various spatial arrangements, significantly improving user experience by ensuring seamless interaction and immediate visual feedback.

5 Conclusions

In this work, we introduced Hybrelighter, a hybrid scene reconstruction and mesh-based filtering approach leveraging guided deep anisotropic diffusion for real-time relighting on edge devices. Our approach successfully combines the computational efficiency of 2D image filters with the depth and geometric accuracy achievable from scene reconstruction capabilities commonly integrated into mixed reality hardware. To address visual artifacts resulting from low-fidelity meshes typically generated by these devices, we demonstrated how guided deep anisotropic diffusion can effectively refine relighting outcomes. Furthermore, we validated the transferability of anisotropic diffusion



Screenshot of our demo app

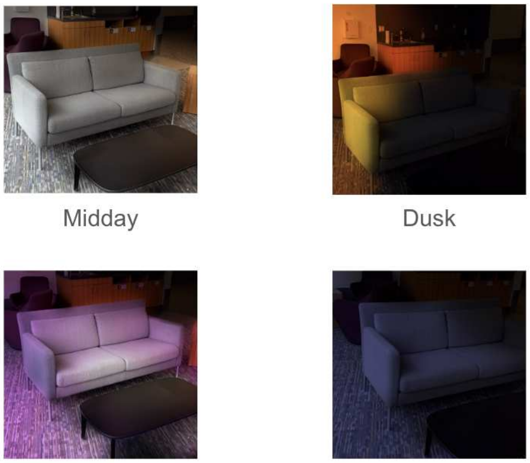
Current time

Early in the morning

Sun fully comes out

Dusk time

Fig. 12: Our demo running in real-time at 50 fps on an iPhone 16 Pro, scanning the environment. We map the positions of the real windows into our demo in order to visualize the lighting more accurately. The top-right image shows what the actual environment looks like. The bottom three images show different times of the day. All images are captured at the same time of the day.



Midday

Dusk

Custom Lightings

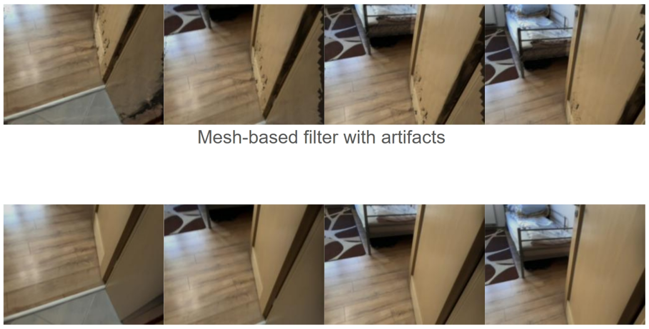
Midnight

Fig. 13: Examples of the room tour demo visualizing different times of the day, and manually placing lights around the room.

models originally trained for guided depth super-resolution tasks to our scenario with minimal adjustments.

We also proposed various improvements, including cascaded anisotropic diffusion and a dedicated shadow processing pass, significantly boosting performance and preserving critical visual details such as shadows. Experimental evaluations against direct mesh-based methods confirm that our approach produces visually superior results, capturing essential high-frequency image features, like edges, more effectively. Additionally, comparisons against high-fidelity dense point cloud renderings underline our method's accuracy and realism.

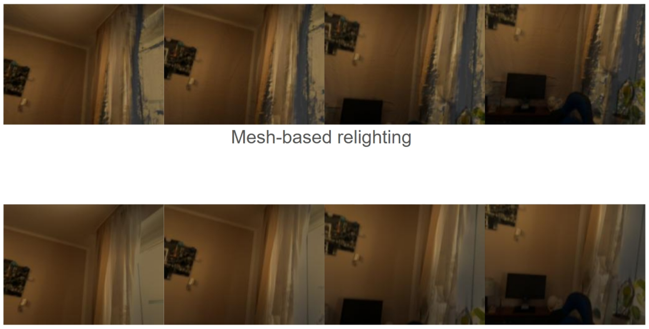
Looking ahead, several promising avenues could further enhance our approach. A key limitation identified in our current pipeline is the absence of a unified framework capable of consistently distinguishing between shadows and geometry-related errors introduced by scanning inaccuracies. Resolving this ambiguity remains a challenging, ill-posed problem. Future research could explore additional learning-based strategies, as well as complementary learning-free methods, to reliably disentangle shadow regions from reconstruction artifacts, ultimately leading to even more robust and visually consistent real-time relighting solutions for mixed reality applications.



Mesh-based filter with artifacts

Our refined output

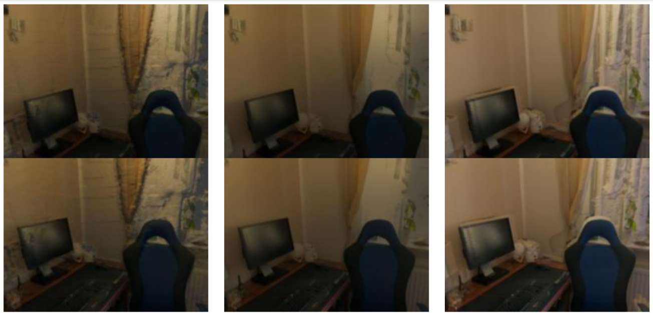
Fig. 14: Side by side comparison showing how anisotropic diffusion can correct the artifacts that exist in the mesh directly acquired by ARKit.



Mesh-based relighting

Our refined output

Fig. 15: Side by side comparison showing our method and direct mesh-based filter approach on relighting tasks.



Mesh-based

Our refined

Dense point clouds

Fig. 16: We compare the direct mesh-based filter approach, our method, and a filter based on high-fidelity point clouds.

Figure Credits

Figure 9 is a creation from [17]

Acknowledgments

The authors wish to thank A, B, and C. This work was supported in part by a grant from XYZ (# 12345-67890).

References

[1] G. Baruch, Z. Chen, A. Dehghan, T. Dimry, Y. Feigin, P. Fu, T. Gebauer, B. Joffe, D. Kurz, A. Schwartz, and E. Shulman. ARKitscenes - a diverse real-world dataset for 3d indoor scene understanding using mobile RGB-d data. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021. 5

[2] A. Bhattad, J. Soole, and D. Forsyth. Stylitgan: Image-based relighting via latent control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[3] Z. Bi, Y. Zeng, C. Zeng, F. Pei, X. Feng, K. Zhou, and H. Wu. Gs^3: Efficient relighting with triple gaussian splatting. In *SIGGRAPH Asia 2024 Conference Papers*, 2024. 2

[4] S. Chen, H. Guo, S. Zhu, F. Zhang, Z. Huang, J. Feng, and B. Kang. Video depth anything: Consistent depth estimation for super-long videos. *arXiv:2501.12375*, 2025. 2

[5] M. V. Conde, F.-A. Vasluianu, J. Xiong, W. Ye, R. Ranjan, and R. Timofte. Compressed depth map super-resolution and restoration: Aim 2024 challenge results, 2024. 3

[6] C. Feng, Z. Chen, C. Zhang, W. Hu, B. Li, and L. Ge. Real-time monocular depth estimation on embedded systems. In *IEEE International Conference on Image Processing, ICIP 2024, Abu Dhabi, United Arab Emirates, October 27-30, 2024*, pp. 3464–3470. IEEE, 2024. doi: 10.1109/ICIP51287.2024.10648152 2

[7] J. Gu, Z. Xiang, Y. Ye, and L. Wang. Denselidar: A real-time pseudo dense depth guided depth completion network. *IEEE Robotics and Automation Letters*, 6(2):1808–1815, 2021. doi: 10.1109/LRA.2021.3060396 3

[8] M. Hu, W. Yin, C. Zhang, Z. Cai, X. Long, H. Chen, K. Wang, G. Yu, C. Shen, and S. Shen. Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. 2024. 2

[9] K. Jacobs and C. Loscos. Classification of illumination methods for mixed reality. *Comput. Graph. Forum*, 25:29–51, 03 2006. doi: 10.1111/j.1467-8659.2006.00816.x 2

[10] H. Jin, Y. Li, F. Luan, Y. Xiangli, S. Bi, K. Zhang, Z. Xu, J. Sun, and N. Snavely. Neural gaffer: Relighting any object via diffusion. In *Advances in Neural Information Processing Systems*, 2024. 2

[11] H. Kim, M. Jang, W. Yoon, J. Lee, D. Na, and S. Woo. Switchlight: Co-design of physics-driven architecture and pre-training framework for human portrait relighting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25096–25106, June 2024. 2

[12] Y. Li, Q. Zhou, J. Gong, Y. Zhu, R. Dazeley, X. Zhao, and X. Lu. Dapointr: Domain adaptive point transformer for point cloud completion, 2024. 3

[13] M. Liu, C. Zeng, X. Wei, R. Shi, L. Chen, C. Xu, M. Zhang, Z. Wang, X. Zhang, I. Liu, H. Wu, and H. Su. Meshformer: High-quality mesh generation with 3d-guided reconstruction model. *arXiv preprint arXiv:2408.10198*, 2024. 3

[14] S. Liu, J. Chen, and J. Zhu. Hvofusion: Incremental mesh reconstruction using hybrid voxel octree. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pp. 6850–6858. International Joint Conferences on Artificial Intelligence Organization, 2024. 2

[15] Y. Liu, S. Dong, S. Wang, Y. Yin, Y. Yang, Q. Fan, and B. Chen. Slam3r: Real-time dense scene reconstruction from monocular rgb videos. *arXiv preprint arXiv:2412.09401*, 2024. 2

[16] Y. Liu, H. Guan, C. Luo, L. Fan, J. Peng, and Z. Zhang. Citygaussian: Real-time high-quality large-scale scene rendering with gaussians, 2024. 2

[17] N. Metzger, R. C. Daudt, and K. Schindler. Guided depth super-resolution by deep anisotropic diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18237–18246, June 2023. 3, 5, 6, 8

[18] P. Ponglertnapakorn, N. Tritrong, and S. Suwajanakorn. Difareli: Diffusion face relighting. 2023. 2

[19] Y. Shi, M. K. Singh, H. Cai, and F. Porikli. Decotr: Enhancing depth completion with 2d and 3d attentions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10736–10746, June 2024. 3

[20] Z. Sun, W. Ye, J. Xiong, G. Choe, J. Wang, S. Su, and R. Ranjan. Consistent direct time-of-flight video depth super-resolution. *arXiv preprint arXiv:2211.08658*, 2022. 3

[21] Y. Wang, G. Zhang, S. Wang, B. Li, Q. Liu, L. Hui, and Y. Dai. Improving depth completion via depth feature upsampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21104–21113, 2024. 3

[22] X. Wei, K. Zhang, S. Bi, H. Tan, F. Luan, V. Deschaintre, K. Sunkavalli, H. Su, and Z. Xu. Meshlrm: Large reconstruction model for high-quality mesh. *arXiv preprint arXiv:2404.12385*, 2024. 2

[23] H.-H. Yang, W.-T. Chen, H.-L. Luo, and S.-Y. Kuo. Multi-modal bifurcated network for depth guided image relighting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2021. 2

[24] R. Yi, C. Zhu, and K. Xu. Weakly-supervised single-view image relighting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8402–8411, June 2023. 2

[25] X. Y. Yuda Song, Zehao Sun. Sdxs: Real-time one-step latent diffusion models with image conditions. *arxiv*, 2024. 2

[26] C. Zeng, G. Chen, Y. Dong, P. Peers, H. Wu, and X. Tong. Relighting neural radiance fields with shadow and highlight hints. In *ACM SIGGRAPH 2023 Conference Proceedings*, 2023. 2

[27] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 5

[28] Z. Zhao, J. Zhang, X. Gu, C. Tan, S. Xu, Y. Zhang, R. Timofte, and L. V. Gool. Spherical space feature decomposition for guided depth map super-resolution, 2023. 3