Atmosphere-Aware Intrinsic Decomposition from a Single Terrain Image with Latent Diffusion Models

Shun Tatsukawa[1] Syuhei Sato[1,2]

[1]Hosei University, Japan [2]Prometech CG Research, Japan

Abstract

We present a latent diffusion-based framework for atmosphere-aware intrinsic decomposition from a single terrain image. In large-scale terrain scenes, atmospheric effects such as fog, haze, and aerial perspective are strongly entangled with surface reflectance and shading, making intrinsic decomposition highly ill-posed. Existing deep learning-based methods mainly focus on albedo and shading estimation and do not explicitly model atmospheric scattering, which often leads to degraded decomposition quality in outdoor scenes with large depth variations. To address this problem, we propose a terrain-specific decomposition framework that explicitly separates an input image into albedo, shading, and atmospheric volume components. We construct a large-scale synthetic terrain dataset with ground-truth intrinsic layers and train a latent diffusion model to estimate them from a single image. To improve compositing consistency, our method incorporates constraints based on a renderer-pass-inspired compositing equation during training and reconstruction guidance during inference. Experimental results demonstrate that the proposed method enables more reliable decomposition of terrain images and supports atmosphere-aware image editing.

Keywords: Intrinsic image decomposition, atmospheric scattering, latent diffusion models, terrain images, image editing

CCS Concepts

• **Computing methodologies** → *Image manipulation; Rendering;*

1. Introduction

For physically accurate and intuitive image editing, numerous advanced image editing techniques, such as retexturing and relighting, have recently been studied. To achieve such editing, it is important to separate diffuse reflectance and shading from an input image, and many deep learning-based intrinsic image decomposition and inverse rendering methods for outdoor scenes have been proposed [BDL*21, CA23]. Many of these conventional methods simplify the problem by decomposing an input image into albedo and a single grayscale shading component; however, this assumption does not account for colored illumination, interreflections, and specular reflections, which can lead to inaccurate reflectance estimation and unnatural color changes during editing. In contrast, Careaga et al. [CA24] proposed a decomposition model into albedo A, color diffuse shading D, and a non-diffuse residual R ($I = A \times D + R$), enabling the separation of more complex reflection components.

However, many conventional methods, including the above, work well for man-made objects such as buildings and roads, but their accuracy often degrades in large-scale natural terrain scenes such as mountainous landscapes. The main reason is that these methods focus only on surface reflection properties and do not explicitly account for light scattering by participating media in the atmosphere, namely, atmospheric volume effects such as fog and haze. Therefore, in natural terrain scenes with large depth variations, contrast reduction and color shifts in distant regions caused by atmospheric scattering are mixed into albedo and shading, making it difficult to correctly estimate surface reflectance and shading components.

To address these issues, we propose an intrinsic image decomposition method for large-scale natural terrain scenes that explicitly separates a volume component representing scattering by atmospheric media, in addition to albedo and diffuse/specular shading. Our method leverages a synthetic terrain dataset based on an atmospheric scattering model and the strong generative capability of diffusion models to accurately estimate intrinsic components. Furthermore, we introduce training with losses based on a compositing equation inspired by the render passes of Cycles [Ble26] and inference guidance that minimizes the reconstruction residual, thereby constraining the estimated components to reconstruct the input image under this compositing model. As a result, albedo, shading, and atmospheric scattering components are decomposed while maintaining compositing consistency, enabling reliable decomposition even for terrain images containing fog and water regions. Furthermore, by learning how changes in the density parameters of atmospheric components, such as air molecules and aerosol particles, multiplicatively affect the

decomposed components, our method enables advanced image editing that reproduces user-specified atmospheric conditions, such as fog density and light scattering. Our source code and dataset are publicly available at `https://sttkw.github.io/terrain-atmospheric-Intrinsic-decomposition/`.

2. Related Work

2.1. Intrinsic Image Decomposition

Many deep learning-based intrinsic image decomposition and inverse rendering methods for outdoor scenes have been proposed [BDL*21, DKG22, CA23, CA24, XPY*24]. Many early methods simplified the problem by modeling an input image as the product of albedo and a single grayscale shading component. ShadingNet [BDL*21] introduced a model that decomposes direct illumination, indirect illumination, and shadows in detail, aiming to improve shading estimation for outdoor images. PIE-Net [DKG22] targets urban scenes and road environments, enabling intrinsic component estimation suited to real-world outdoor scenes. Furthermore, the method of Careaga et al. [CA23] focuses on ordinal relationships between pixels and demonstrates stable decomposition performance even for high-resolution real-world outdoor images. However, all of these methods are essentially based on decomposition into albedo and grayscale shading, making it difficult to explicitly handle colored illumination, interreflections, and specular reflections.

To address this limitation, Careaga et al. [CA24] proposed an analytical modeling approach that decomposes an image into albedo A, color diffuse shading D, and a non-diffuse residual R representing effects such as specular reflection ($I = A \times D + R$). This enables more appropriate separation of colored illumination and interreflections, while also supporting advanced image editing operations such as specular highlight removal and local white-balance correction. Similarly, Chen et al. [XPY*24] proposed a method that learns the distributions of albedo A and specular shading S using a diffusion model, and uses them as guidance for differentiable rendering to estimate surface roughness and apply the results to relighting.

However, many existing methods, including these approaches, still mainly focus on reflection properties at object surfaces, namely the interaction between material and illumination. Therefore, light transport through participating media in space, that is, atmospheric volume effects such as fog and haze, which are prominent in mountainous landscapes and large-scale natural terrain scenes, is not explicitly considered. As a result, in vast natural terrain scenes, contrast reduction and color shifts in distant regions may be incorrectly mixed into albedo and shading.

2.2. Diffusion Models for Inverse Rendering and Decomposition

In recent years, methods that apply powerful diffusion models to inverse rendering and estimate various G-buffers, such as material and geometric properties, from a single image in the "$RGB \leftrightarrow X$" setting have attracted increasing attention. For example, Zeng et al. [ZDG*24] and Liang et al. [LGL*25] exploit the strong priors of diffusion models and achieve a certain degree of generalization to complex real-world scenes and outdoor environments, even though they are mainly trained on synthetic indoor-scene data.

However, while these methods primarily focus on material estimation, intrinsic image decomposition for image editing faces a different challenge: the generative nature of diffusion models can compromise fidelity to the input image [CA24]. In addition, when each intrinsic component is estimated independently and the image is reconstructed from them, brightness and color inconsistencies can arise between the reconstructed image and the original input image [ZDG*24, LGL*25]. Therefore, to obtain intrinsic components that can be reliably used for image editing while maintaining the generalization ability of diffusion models, a new framework is required that achieves both image fidelity and consistency between the reconstructed and input images.

2.3. Synthetic Datasets for Vision and Graphics

The performance of the above deep learning-based models strongly depends on the quality and quantity of the training data. Many datasets that provide accurate ground truth [LYS*21, RRR*21, ZLH*22] are synthetic datasets designed for indoor environments. By combining them with a small amount of real-world data [GJAF09, BBS14], existing methods can decompose unseen real-world outdoor images with a certain level of accuracy. However, the data distributions learned by these methods are still centered on indoor scenes and nearby objects. Therefore, when they are applied to large-scale natural landscapes such as mountainous terrain, they fail to achieve accurate decomposition, particularly for water regions and hazy distant views.

On the other hand, several synthetic datasets have also been constructed specifically for outdoor scenes [BDL*21, LDM*21]. Although these datasets are useful for learning complex outdoor geometry and textures, one issue is that they rely on environment maps such as high-dynamic-range images (HDRIs) to represent scene illumination. HDRIs are effective for capturing skylight and sunlight from infinitely distant directions, but they cannot model the effects of participating media existing in the 3D space from the light source or objects to the camera. In large-scale natural landscapes such as mountainous terrain, the atmosphere itself behaves as a huge volume and causes complex optical phenomena such as Rayleigh scattering and Mie scattering, depending on the distance light travels through the medium and its physical density. Approaches based on 2D environment maps cannot explicitly parameterize such spatially extended atmospheric volumes. As a result, it is difficult to achieve image editing that physically and naturally manipulates atmospheric conditions such as fog and haze.

3. Background Models

Our image decomposition and atmospheric editing pipeline, as well as the dataset used to train these components, are based on several existing methods. We describe their details below.

3.1. InstructPix2Pix

Wang et al. [WZZ*22] showed that, for image translation tasks with limited paired training data, fine-tuning a large-scale image diffu-

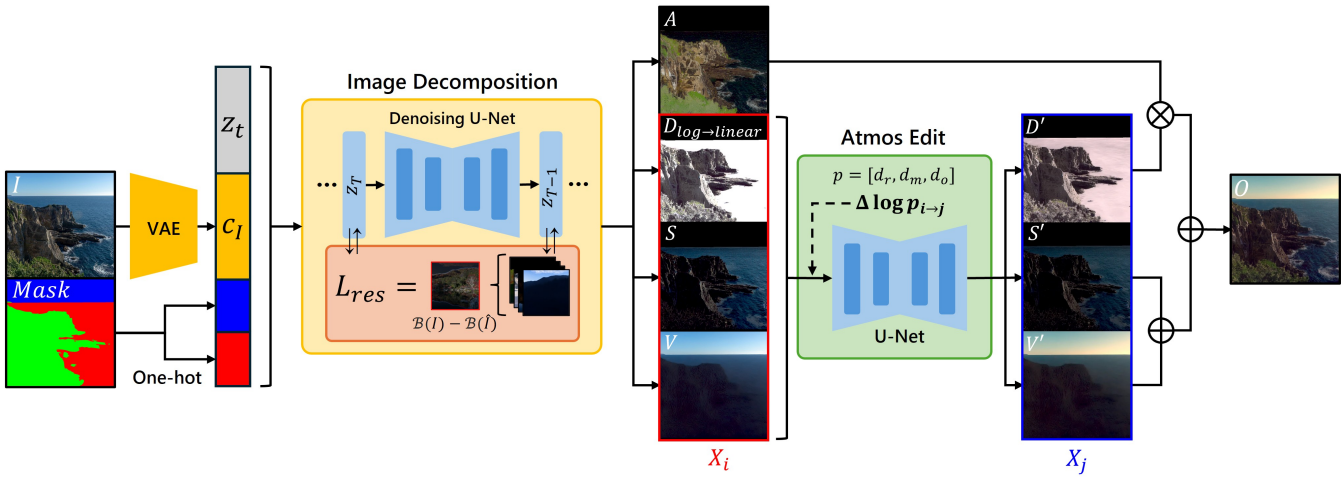


Figure 1: *Overview of our framework.*

sion model is more effective than training a model from scratch. Therefore, we adopt InstructPix2Pix by Brooks et al. [BHE23] as the foundation of our image decomposition model and fine-tune it for our task. InstructPix2Pix is a diffusion-based image translation model that edits an input image according to a text instruction by combining the knowledge of two large-scale models pretrained on different modalities. The weights of this model are initialized from the pretrained Stable Diffusion-1.5 [RBL*22], and the model is trained to generate the latent representation of the target image conditioned on the input image c_I and text instruction c_T, in addition to the latent noise z_t. However, while the original InstructPix2Pix uses noise prediction, we adopt v-prediction, i.e., velocity prediction, as the objective in our fine-tuning. This is because, in tasks that estimate physical parameters or intrinsic components from images, using v-prediction has been reported to stabilize training and empirically produce higher-quality results [ZDG*24].

3.2. Atmospheric Scattering

In our method, we simulate environmental illumination and atmospheric scattering in the terrain dataset using the physically based atmospheric rendering method proposed by Hillaire et al. [Hil20]. This method accurately simulates radiative transfer in the media that constitute the atmosphere, and it represents atmospheric effects mainly through three physical components. The first is Rayleigh scattering by air molecules, the second is Mie scattering by aerosol particles such as dust and pollutants in the atmosphere, and the third is absorption of specific wavelengths of light by ozone molecules. Let d_r, d_m, and d_o denote the density scale coefficients of air molecules, aerosol particles, and ozone molecules, respectively. Then, the scattering coefficient σ_s and absorption coefficient σ_a are expressed as linear combinations of the reference physical coefficients of each medium $(\sigma_s^r, \sigma_s^m, \sigma_a^m, \sigma_a^o)$ and the corresponding density scale coefficients as follows.

$$\sigma_s = d_r \sigma_s^r + d_m \sigma_s^m \tag{1}$$

$$\sigma_a = d_m \sigma_a^m + d_o \sigma_a^o \tag{2}$$

The extinction coefficient σ_t, which represents the total attenuation of light traveling through the medium, is defined as the sum of the scattering and absorption coefficients.

$$\sigma_t = \sigma_s + \sigma_a. \tag{3}$$

These three density scale coefficients change the wavelength-dependent scattering and absorption properties of light, and appear in the image as color variations characteristic of atmospheric volumes, such as the blueness of the sky, the redness of sunsets, and the whiteness of fog. The method of Hillaire et al. and its physical coefficients are implemented as a module using Geometry Nodes in Blender [Ble26] [Wol]. By changing the scalar density coefficients d_r, d_m, and d_o, the appearance of the sky and atmospheric scattering can be controlled. We use this module to control the three density scale coefficients and generate a dataset for image decomposition under diverse atmospheric conditions. Furthermore, in atmospheric editing, we use relative changes in these density parameters as conditions to learn corresponding changes in the intrinsic component space. This enables users to specify relative changes in the density parameters and reflect the resulting changes in fog density and light scattering through updates to each intrinsic component.

4. Proposed Method

The overall pipeline of our method is shown in Fig. 1. The latent representation c_I and the input image I, mapped into the latent space by a VAE, are provided to the image decomposition model conditioned on the mask information. In the image decomposition phase, albedo A, diffuse shading D, specular shading S, and volume V are estimated individually according to text instructions, and a correction based on the reconstruction error is applied during the denoising process. In the atmospheric editing phase, D, S, and V are treated as the editable components. The log-difference $\Delta \log p_{i \to j}$ is computed from the current atmospheric parameters $p_i = [d_r, d_m, d_o]$ and the target atmospheric parameters p_j, and is used as a condition for the U-Net to predict the change for each

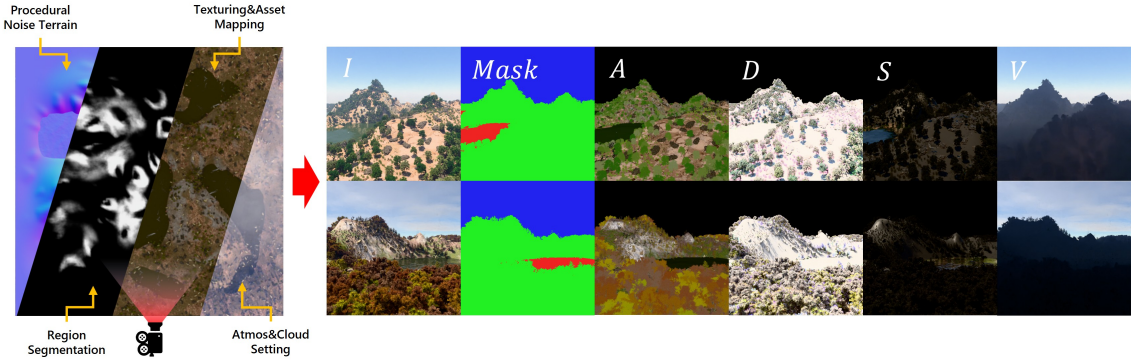


Figure 2: *Overview of the dataset generation pipeline and data structure.*

component. Based on the predicted changes, D, S, and V are updated, and the edited components are composited according to the renderer-pass-inspired compositing equation to obtain the output image O with modified atmospheric conditions.

4.1. Intrinsic Components

In our method, image formation is modeled not as a complete radiative transfer equation but as the following compositing equation inspired by the render passes of Cycles [Ble26]:

$$I = A \cdot D + S + V \tag{4}$$

Here, I represents an LDR linear image in the range of $[0,1]$, as in the input image. Each intrinsic component has the following physical meaning:

- **Albedo** A: Represents the terrain-specific color and diffuse reflectance.
- **Diffuse Shading** D: Represents the intensity and color of diffuse shading caused by terrain shadows and interreflections, with atmospheric attenuation along the surface-to-camera path already incorporated.
- **Specular Shading** S: Represents the intensity and color of specular shading, capturing reflections of the sky and terrain on water surfaces, with atmospheric attenuation along the surface-to-camera path already incorporated.
- **Volume** V: Represents the additive scattered light accumulated along camera rays through volumetric media such as the atmosphere and fog.

In Cycles, the Diffuse/Glossy Direct and Indirect passes used for D and S include atmospheric attenuation. Thus, Eq. (4) separates attenuated surface contributions from the additive in-scattered radiance V, rather than explicitly recovering transmittance.

During training, albedo A is treated as an LDR component in the range of $0 \sim 1$, corresponding to material color. The goal of this study is to enable image decomposition and atmospheric editing in the same LDR image space as the input image, rather than to fully recover HDR radiance values. Therefore, the additive components, specular shading S and volume V, are directly learned by clipping their linear values to $0 \sim 1$ as LDR representations for image editing. In contrast, diffuse shading D is treated as an HDR (High Dynamic Range) component because it acts as a multiplicative component on albedo and needs to represent strong illumination and highlights with values greater than 1. Careaga et al. [CA24] pointed out that strong reflections and extremely large values in diffuse shading form a long-tailed distribution, making direct regression unstable, and therefore used an inverse-space transformation $1/(X+1)$ to map the values into the range of $0 \sim 1$. However, when this inverse-space transformation is applied within a latent diffusion framework, extremely bright highlight regions are compressed into very small differences near 0. As a result, fine gradation information is lost during VAE-based encoding and decoding into the latent space and during the noise addition process, causing highlights to be completely crushed during reconstruction.

Therefore, for diffuse shading D, we transform it into the following logarithmic representation during training to compress its value range according to the representational constraints of the model while safely preserving highlight gradations even in the latent space:

$$D_{\log} = \mathrm{clip}\left(\frac{\log(1+D)}{\log(1+D_{\max})}, 0, 1\right). \tag{5}$$

Here, since the mean of the 99.9th percentile values of the luminance distribution in the ground-truth data was 4.81, we set $D_{\max} = 5$ and ignore values larger than this as outliers. In addition, since typical input images used during inference are stored in the sRGB color space, we convert each pixel to the linear-space image I_{linear} by raising it to the power of 2.2 before feeding it into the model.

4.2. Datasets

Our terrain generation system is implemented in Blender based on the system and asset collection provided by True-VFX [Tru]. We construct a procedural pipeline for generating a complex and photorealistic terrain dataset so that the diffusion model can generalize across diverse materials and atmospheric conditions. All scenes are rendered using Cycles at a resolution of 512×512.

An overview of our dataset generation pipeline is shown in Fig. 2. First, terrain meshes are generated using fractional Brownian motion (fBm) noise derived from integer seeds ranging from 1 to 4,000 as height maps. The generated terrain is divided into cliff

and flat regions using a gradient threshold of 0.20, and each region is further divided into two using fractal noise with a boundary width of 0.25, resulting in four material regions in total. For each material region, one texture is independently selected from 166 terrain texture candidates according to a discrete uniform distribution.

Furthermore, we place a total of ten types of scattered assets, consisting of three types of bushes, two types of dead trees, two types of grass, and three types of rocks. For each category, whether the asset is placed is independently determined according to a Bernoulli distribution. When present, its placement density and scale are independently sampled from uniform distributions and assigned to the corresponding scattering parameters. We also introduce a ten-level discrete seasonal parameter for vegetation assets, with one level uniformly sampled for each scene. The selected seasonal level determines common hue and saturation transformations that are applied consistently to all vegetation assets in the scene, allowing their appearance to vary coherently from green foliage to autumn colors. This provides a more diverse and realistic distribution of vegetation appearances. Furthermore, we place a water-region mesh and independently sample the wave scale, the amount of foam generated according to surface undulations, and a water-quality parameter that determines the base color of the water material from uniform distributions. The colors and patterns originating from the water surface geometry and water quality are included in Albedo A, whereas reflections caused by illumination and the surrounding environment are separated into Specular Shading S. This design allows the model to distinguish the intrinsic appearance of the water surface from reflected light.

The ground-truth intrinsic components are generated from the corresponding Cycles render passes. The LDR clipping and logarithmic transformation described in Section 4.1 are then applied to the respective components. The illumination environment and atmospheric appearance of the scenes constructed in this way are represented using the physically based atmospheric rendering method of Hillaire et al. [Hil20], as described in Section 3.2. The density scale coefficients of air molecules, aerosol particles, and ozone molecules that constitute atmospheric scattering, as well as the angle of sunlight, are uniformly sampled from the following ranges to avoid confusing the model during training with extreme parameters:

$$\begin{aligned} \Theta &\sim \mathcal{U}(10,90) \quad [\text{deg}], \\ \Phi &\sim \mathcal{U}(0,360) \quad [\text{deg}], \\ d_r, d_m, d_o &\sim \mathcal{U}(0.3,5.0) \end{aligned}$$

Here, Θ denotes the solar elevation angle, Φ denotes the azimuth angle, and d_r, d_m, d_o are defined as dimensionless scalar quantities representing the density scales of the respective atmospheric components. We construct two datasets with different data structures according to the characteristics of the learning task required for each module. The first dataset is designed for the image decomposition module. For terrain scenes generated from 4,000 different seed values, we render each scene from four camera angles that include different viewing distances from near to far, resulting in a total of 16,000 image pairs. The second dataset is designed for the atmospheric editing module. For terrain scenes generated from 1,000 different seed values, we fix the camera position and terrain geometry while randomly varying only the atmospheric parameters to render 10 image patterns, resulting in a total of 10,000 data samples. This design enables the network to learn how relative changes in the density parameters (d_r, d_m, d_o) affect D, S, and V under fixed terrain geometry.

4.3. Intrinsic Image Decomposition

For the image decomposition task, we fine-tune a model based on InstructPix2Pix [BHE23], as described in Section 3.1. In our framework, the type of output is specified by providing one of the prompts, "Albedo," "Diffuse Shading," "Specular Shading," or "Volume," as the text condition c_T, and the model then infers the corresponding intrinsic map. Attempting to predict all components simultaneously by expanding the output channels of the diffusion model would require retraining the input and output convolution layers from scratch, which significantly degrades generation quality [ZDG*24]. Therefore, we adopt this design.

Furthermore, to improve decomposition accuracy in complex natural scenes, we concatenate a two-class segmentation map of water and sky regions corresponding to the input image as one-hot channels, together with the latent noise z_t and the VAE-encoded input image c_I, and feed them into the network. Combined with the water-region loss described later, this enables the model to explicitly utilize region semantics at each pixel, thereby strongly suppressing erroneous leakage between strong specular reflections on water surfaces and volumetric effects in the sky. For synthetic images, the water and sky masks are obtained from the ground-truth render passes. For real images, the masks used in the experiments are obtained semi-automatically using image-editing software, with a rough automatic selection followed by manual correction, requiring approximately five minutes per image. Since boundaries of water and sky regions in natural terrain images can be ambiguous, and errors in the masks directly affect the decomposition results, fully automatic mask generation using a semantic segmentation model is left for future work.

Training is performed using images with a resolution of 512×512, a learning rate of 1.0×10^{-5}, a batch size of 128, and 30 epochs. During inference, to prevent overexposure of predicted pixel values, which can severely degrade intrinsic map estimation, we adopt a DDIM scheduler based on the findings of Lin et al. [LLLY24].

4.3.1. Loss Functions

To balance the high-quality image synthesis capability of the generative model with compositing consistency based on the compositing equation in Eq. (4), we minimize the weighted sum of the following four loss functions, denoted as $\mathcal{L}_{\text{decomp}}$, during image decomposition training.

$$\mathcal{L}_{\text{decomp}} = \mathcal{L}_{\text{diff}} + 0.5\mathcal{L}_{\text{direct}} + \mathcal{L}_{\text{rec}} + 2.0\mathcal{L}_{\text{water}}. \tag{6}$$

Throughout this section, for each intrinsic component X, $\hat{X}$ and X^* denote the predicted and ground-truth values, respectively. $\mathcal{L}_{\text{diff}}$ is the diffusion loss. The input image I and region mask M are provided as conditions, and training is performed for the component type $k \in \{A, D, S, V\}$ and diffusion timestep t. Here, $z_{k,t}$ denotes the

latent representation of component k with noise added at timestep t, and $v_{k,t}$ is the velocity representation defined as a linear combination of the clean latent representation and the noise component. This loss trains the model to predict the velocity for each intrinsic component by minimizing the squared error between the predicted velocity $\hat{v}_\theta$ and the target velocity $v_{k,t}$ over component types and diffusion timesteps.

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{k,t}\left[\left\|\hat{v}_\theta(z_{k,t}, I, M, k, t) - v_{k,t}\right\|_2^2\right] \tag{7}$$

$\mathcal{L}_{\text{direct}}$ is the direct loss, which computes the L1 error between each component $\hat{X}$ estimated by the network and its corresponding ground truth X^*. A, S, and V are defined in linear space, while D is defined in the space of the range-compressed logarithmic representation D_{log}.

$$\mathcal{L}_{\text{direct}} = \sum_{X \in \{A, D_{\log}, S, V\}} \left\|\hat{X} - X^*\right\|_1. \tag{8}$$

$\mathcal{L}$rec is the reconstruction loss. For each scene, the latent inputs corresponding to the four component prompts are stacked along the batch dimension and estimated simultaneously by the shared U-Net. An image is composited in linear space from the estimated components using Eq. (4) ($\hat{I} = \hat{A} \cdot \hat{D} + \hat{S} + \hat{V}$), clamped to the range of $0 \sim 1$, and its L1 error from the input image converted into linear RGB space, Ilinear, is computed.

$$\mathcal{L}_{\text{rec}} = \left\|\text{clamp}(\hat{I}, 0, 1) - I_{\text{linear}}\right\|_1 \tag{9}$$

This loss is backpropagated through all four component prediction paths. Finally, $\mathcal{L}_{\text{water}}$ is the water-region mask loss. It smooths the spatial gradients of albedo, diffuse shading (D_{log}), and volume within the water-region mask, while strongly assigning local reflection structures to specular shading. Here, $\text{mean}_{\text{water}}(\cdot)$ denotes the average within the water-region mask.

$$\mathcal{L}_{\text{water}} = \sum_{X \in \{A, D_{\log}, V\}} \text{mean}_{\text{water}}\left(|\nabla \hat{X}|\right) + \text{mean}_{\text{water}}\left(|\hat{S} - S^*|\right). \tag{10}$$

The image-space losses ($\mathcal{L}_{\text{direct}}, \mathcal{L}_{\text{rec}}, \mathcal{L}_{\text{water}}$) are weighted according to the diffusion timestep t using $w(t) = \sqrt{\bar{\alpha}_t}$. $\bar{\alpha}_t$ increases as denoising progresses and the image becomes clearer. This prevents training instability caused by enforcing constraints in the early stages of generation, where noise is dominant, and the structures of the components are not yet determined.

4.4. Inference with Reconstruction Guidance

During inference on an unknown image, independently generating each component may cause discrepancies in brightness and color between the image reconstructed in linear space and the input image I. On the other hand, in our framework based on a pretrained model, it is difficult to achieve decomposition that always guarantees perfect reconstruction, because all components are not decomposed simultaneously.

To address this issue, during inference, we regard the distribution represented by the trained diffusion model as a prior that enforces the naturalness of each component, and perform sampling from a posterior distribution in which the compositing equation in Eq. (4) serves as a weak observation constraint. Specifically, we compute the reconstructed image $\hat{I} = \hat{A} \cdot \hat{D} + \hat{S} + \hat{V}$ from the current estimated components, and define the reconstruction residual as the difference from the input image I in a low-frequency image space obtained by applying a smoothing operator $\mathcal{B}(\cdot)$:

$$\mathcal{L}_{\text{res}} = \left\|\mathcal{B}(\hat{I}) - \mathcal{B}(I)\right\|_1 \tag{11}$$

Here, $\mathcal{B}$ applies average pooling with a kernel size and stride of 8, followed by bilinear interpolation to the original resolution. The four component latents are jointly updated over 50 DDIM steps. During the final 15 steps, the latent variables are updated using the normalized gradient of the reconstruction residual:

$$z_t \leftarrow z_t - \eta_t \frac{\nabla_{z_t} \mathcal{L}_{\text{res}}\left(\hat{X}_0(z_t), I\right)}{\left\|\nabla_{z_t} \mathcal{L}_{\text{res}}\left(\hat{X}_0(z_t), I\right)\right\| + \epsilon} \tag{12}$$

Here, the reconstruction residual is computed in a smoothed low-frequency image space, and the update direction is normalized. Therefore, this guidance does not directly force the high-frequency details of the input image, but instead acts as a weak correction signal for global color and brightness consistency. As a result, detailed structures of terrain and reflections are left to the prior distribution of the diffusion model, while only global inconsistencies in color and brightness are corrected. The correction coefficient η_t is a timestep-dependent guidance strength, and ϵ is a small constant for numerical stability. During the final 15 denoising steps, η_t is increased from 0 to 0.02 according to a cosine schedule, thereby safely balancing structure generation in the early noisy stages and compositing consistency in the later stages.

4.5. Atmosphere Editing

In the atmospheric editing phase, we learn how changes in atmospheric conditions affect D, S, and V using a dataset in which only the atmospheric parameters vary under the same terrain and camera conditions. Changes in atmospheric parameters tend to appear not as additive differences in each component, but as multiplicative changes in the strength of scattering and attenuation. Therefore, we represent the difference between the current atmospheric condition p_i and the target condition p_j as a logarithmic difference, $\Delta \log p = \log(p_j + \epsilon_p) - \log(p_i + \epsilon_p)$, where $\epsilon_p = 10^{-6}$ is a small constant added for numerical stability when the parameter values approach zero, and use it as a condition to predict a logarithmic multiplicative field $\widehat{\Delta \log X}$ for each component $X \in \{D, S, V\}$. The three-component vector $\Delta \log p$ is spatially broadcast and concatenated with the current components $[D, S, V]$ along the channel dimension at the U-Net input.

During training, we use $\Delta \log p$ computed from known p_i and p_j, together with the corresponding component pair (X_i, X_j), to learn the component changes caused by the parameter variation. In contrast, during inference, the model only takes the relative change $\Delta \log p$ specified by the user as input, and therefore does not require estimating the current absolute atmospheric parameters p_i from an unknown real image. In other words, our method does not recover absolute atmospheric densities, but performs relative atmospheric editing with respect to the current image state.

Since atmospheric changes are spatially smooth, we smooth the predicted logarithmic change $\widehat{\Delta \log X}$ to obtain $\widehat{\Delta \log X}_{\text{low}}$, where

Table 1: Quantitative evaluation of image decomposition performance on EDEN and photorealistic outdoor scenes.

	Albedo				Diffuse Shading				Specular Shading			
	EDEN		Outdoor Scenes		EDEN		Outdoor Scenes		EDEN		Outdoor Scenes	
Method	PSNR↑	LPIPS↓	PSNR↑	LPIPS↓	PSNR↑	LPIPS↓	PSNR↑	LPIPS↓	PSNR↑	LPIPS↓	PSNR↑	LPIPS↓
Zeng [ZDG*24]	15.5	0.31	13.4	0.40	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Careaga [CA24]	17.6	0.25	13.2	0.45	**12.0**	0.37	10.6	0.39	N/A	N/A	N/A	N/A
Chen [XPY*24]	16.7	0.29	15.6	0.46	N/A	N/A	N/A	N/A	24.2	0.14	17.7	0.32
Liang [LGL*25]	17.9	0.25	16.9	0.40	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Sartor [SP25]	16.6	0.31	18.6	0.47	11.9	0.37	10.8	0.46	N/A	N/A	N/A	N/A
Han [HZT*26]	14.9	0.28	13.7	0.40	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Ours	**18.7**	**0.22**	**20.5**	**0.34**	11.9	**0.32**	**13.1**	**0.37**	**28.6**	**0.13**	**26.2**	**0.22**

the subscript low denotes its smoothed low-frequency component. We then exponentiate it and multiply it by the current component X_i to obtain the edited component $\hat{X}_j$.

$$\hat{X}_j = X_i \odot \exp\left(\widehat{\Delta \log X}_{\text{low}}\right) \tag{13}$$

The training loss for the atmospheric editing model consists of a logarithmic difference loss, which encourages the predicted log-change to match the ground-truth log-change, and a component loss, which encourages the edited component to match the target component:

$$L_{\text{edit}} = L_\Delta + L_X \tag{14}$$

The logarithmic difference loss L_Δ constrains the predicted multiplicative field to match the ground-truth multiplicative field. To produce spatially smooth changes that are plausible as atmospheric variations, the comparison is performed in the low-frequency domain as follows:

$$L_\Delta = \left\| \widehat{\Delta \log X}_{\text{low}} - \Delta \log X^*_{\text{low}} \right\|_1 \tag{15}$$

The component loss L_X constrains the edited D, S, and V to be close to the components under the target condition:

$$L_X = \left\| \hat{X}_j - X_j \right\|_1 \tag{16}$$

Training is performed using images at a resolution of 512×512, with a learning rate of 1.0×10^{-4}, a batch size of 24, and 100 epochs.

5. Experiments

In this section, we evaluate the effectiveness of the proposed method from the perspectives of image decomposition and atmospheric condition editing. Note that the intrinsic components shown in each figure are converted from the values output in linear space to gamma space by raising them to the power of $1/2.2$, in order to make visual differences easier to understand. These experiments are conducted using a single NVIDIA H100 GPU. Intrinsic image decomposition with reconstruction guidance takes 6.69 seconds per image on average, including the time required to save all output images, whereas decomposition without reconstruction guidance takes 3.10 seconds per image. In addition, once the intrinsic components have been obtained, atmospheric editing takes only 28.05 ms per edit on average, enabling interactive adjustment of atmospheric conditions at near-real-time speed.

5.1. Evaluation of Intrinsic Image Decomposition

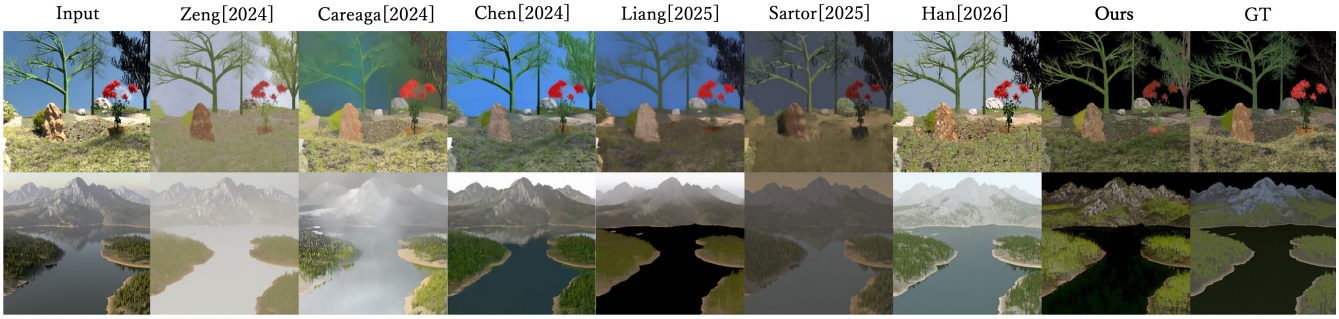
To evaluate the accuracy of image decomposition, we compare our method with conventional methods using synthetic datasets. As the evaluation datasets, we use 125 scenes from EDEN [LDM*21], a multimodal synthetic dataset of outdoor natural scenes, and 15 outdoor evaluation scenes constructed from landscape scenes obtained from BlenderKit [Ble]. The latter consist of a small number of photorealistic scenes manually created by multiple designers and are used as an approximate quantitative evaluation of real-world scenes, for which obtaining ground-truth intrinsic components is difficult.

We compare albedo estimation with six baseline methods: Zeng et al. [ZDG*24], Careaga et al. [CA24], Chen et al. [XPY*24], Liang et al. [LGL*25], Sartor et al. [SP25], and Han et al. [HZT*26]. Among these methods, diffuse shading is additionally compared with Careaga et al. and Sartor et al., and specular shading is compared with Chen et al. These methods have demonstrated generalization to outdoor scenes, and we use their pretrained models for comparison.

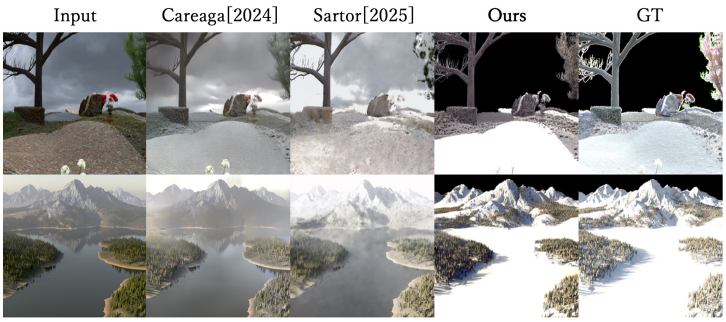
For each method, we follow its original output definition and convert the predicted components to the corresponding linear-RGB space before evaluation. As evaluation metrics, we use PSNR to measure pixel-level fidelity of the generated images and LPIPS [ZIE*18] to measure perceptual quality. Since the treatment of sky regions is not consistent across conventional methods, we exclude sky regions from the evaluation to ensure a fair comparison. As shown in Table 1 and Fig. 3, the proposed method achieves higher decomposition accuracy than existing methods under most conditions. In particular, on the 15-scene photorealistic evaluation set, our method outperforms all applicable baseline methods for A, D, and S.

5.2. Evaluation on Real-world Images

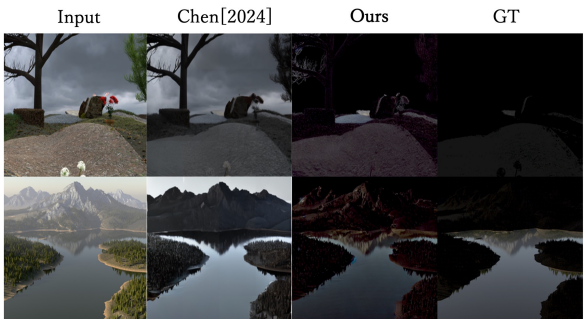
Next, we conducted the same experiment on real-world outdoor images. Since complete ground truth is not available for real images, we perform a visual qualitative evaluation focusing on the effects of fog and water-surface reflections. We compare against the same applicable baseline methods used in the synthetic evaluation. As shown in Fig. 4, existing methods tend to mix water-surface reflections and fog into the albedo, whereas the proposed method can suppress such leakage to some extent. Focusing on water-surface reflections, the method of Careaga et al. [CA24] mixes specular reflections into diffuse shading, while the method of



(a) *Albedo*

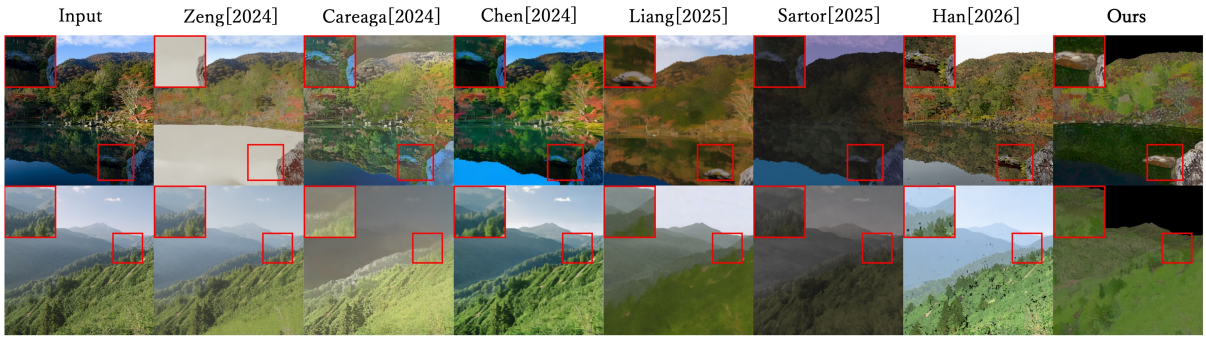


(b) *Diffuse Shading*

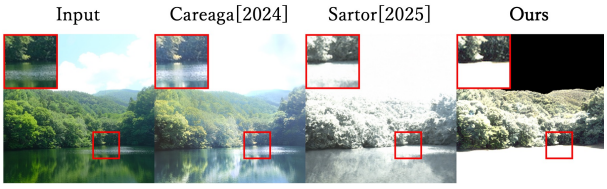


(c) *Specular Shading*

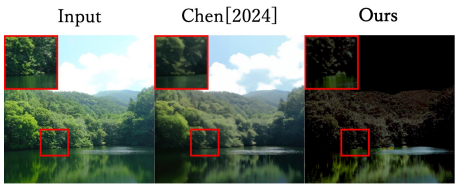
Figure 3: *Image decomposition results on synthetic data from EDEN and photorealistic outdoor scenes.*



(a) *Albedo*



(b) *Diffuse Shading*



(c) *Specular Shading*

Figure 4: *Intrinsic image decomposition results on real images.*

Chen et al. [XPY*24] does not sufficiently capture specular reflections in the specular shading component. In contrast, the proposed method plausibly decomposes diffuse shading and specular shading as their respective reflection components even for real images. Furthermore, Fig. 5 shows the decomposition results of the proposed method on additional real images, together with the reconstruction results based on the compositing equation in Eq. (4). The results show that the intrinsic components are plausibly decomposed for various real images, and that images close to the inputs can be reconstructed.

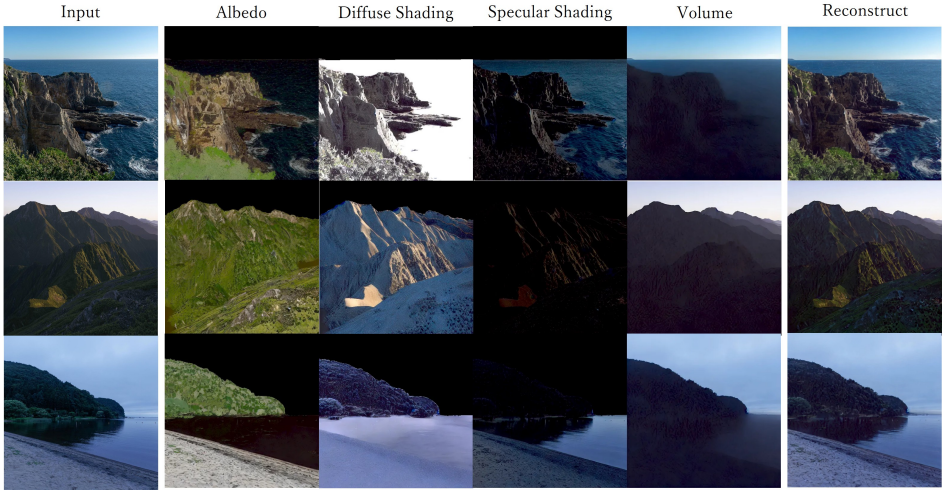


Figure 5: *Additional intrinsic image decomposition results on real images.*

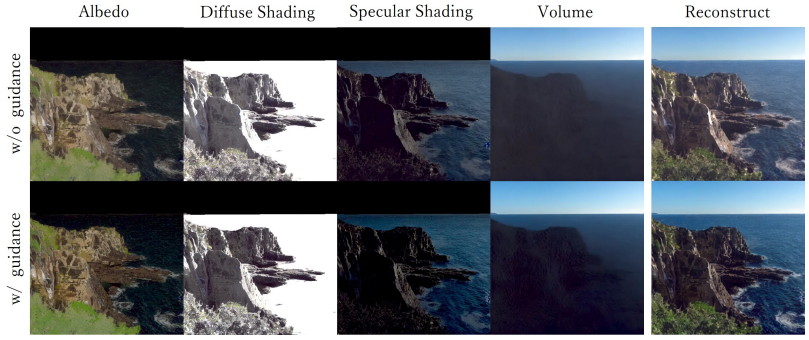


Figure 6: *Ablation study on reconstruction guidance.*

5.3. Ablation Study on Reconstruction Guidance

In the proposed image decomposition network, we add a gradient based on the reconstruction residual as guidance during inference in order to correct slight color discrepancies in the reconstructed image that cannot be sufficiently corrected by the reconstruction loss alone. As shown in Fig. 6, this guidance brings the reconstruction result closer to the input image without causing the residual to concentrate on a specific component. In fact, when comparing the PSNR and LPIPS [ZIE*18] between the input and reconstructed images for 10 real images, we confirmed an improvement in accuracy, as shown in Table 2.

Table 2: Ablation study on reconstruction guidance.

	Recon. Accuracy	
Ablation	PSNR↑	LPIPS↓
w/o guidance	19.2	0.46
w/ guidance	**22.9**	**0.42**

5.4. Ablation Study on Segmentation Mask

In the proposed image decomposition network, we condition the model by concatenating a two-class segmentation map of water and sky regions with the input image to explicitly provide the semantics of terrain regions. To verify the effectiveness of this design, we trained an ablation model in which the segmentation mask was excluded from the input, and compared the decomposition results on real images. As shown in Fig. 7, in the upper-row model without the mask, strong specular reflections on the water surface are mixed into the albedo and diffuse shading components. This result confirms that introducing the segmentation mask is effective for decomposing complex natural landscapes.

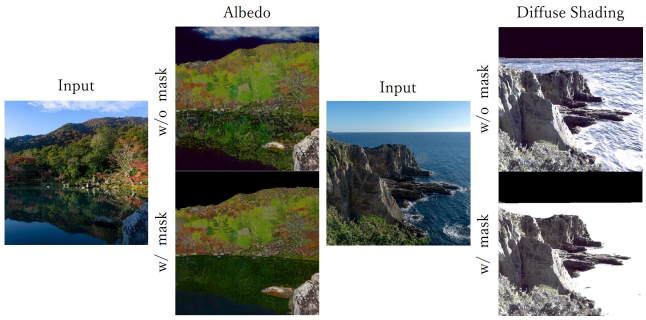


Figure 7: *Ablation study on segmentation mask.*

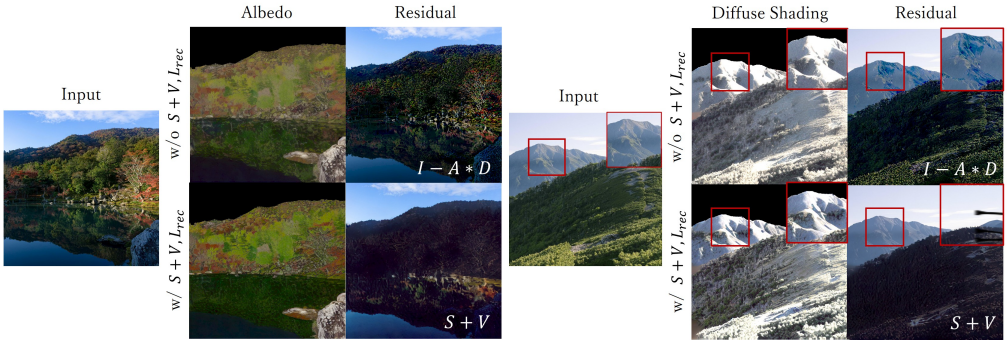


Figure 8: *Ablation study on explicit specular/volume decomposition and reconstruction loss.*

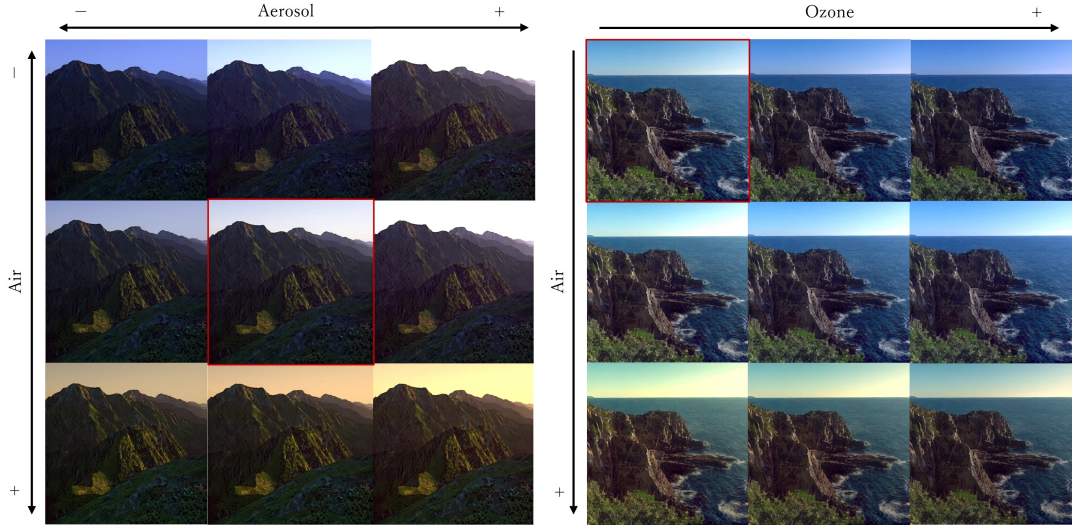


Figure 9: *Editing results of atmospheric conditions. The red box indicates the unedited condition.*

5.5. Ablation Study on Specular Shading and Volume Decomposition

In the proposed image decomposition network, we explicitly decompose specular shading and volume components for effective image decomposition and atmospheric editing of terrain images containing fog and water regions. To verify whether explicitly handling these components contributes to the decomposition accuracy of the other components, we trained an ablation model following the method of Careaga et al. [CA24]. In this model, the compositing equation is formulated as $I = A \cdot D + R$ with a residual R, and only albedo and diffuse shading are inferred by the network without using a reconstruction loss during training. In this experiment, reconstruction guidance during inference is not applied to fairly compare the training conditions. As shown in Fig. 8, the upper-row model, which does not explicitly decompose $S+V$, cannot sufficiently separate specular reflections on water regions and fog from albedo and diffuse shading. In fact, unlike $S+V$ in the proposed method, the residual of the upper-row model clearly contains colors that can be regarded as diffuse components. This confirms that explicitly decomposing specular shading and volume components contributes to improving the decomposition accuracy of the other components.

5.6. Atmosphere Editing

We conducted experiments in which atmospheric parameters were manipulated to reconstruct new images. First, we evaluated the editing results of each component on synthetic data using PSNR and LPIPS [ZIE*18]. As shown in Table 3, our editing model accurately transforms each component. Next, we applied image editing to real-world images. In Fig. 9, the red boxes indicate the unedited reconstruction results, while the other results show edited images obtained by changing each atmospheric parameter. Fig. 10 shows the changes in each component during atmospheric editing. These results demonstrate that our method achieves stable image editing while preserving the terrain details before editing.

Table 3: Editing accuracy of each component on synthetic data.

Diffuse shading		Specular shading		Volume	
PSNR↑	LPIPS↓	PSNR↑	LPIPS↓	PSNR↑	LPIPS↓
26.7	0.040	41.9	0.028	23.7	0.071

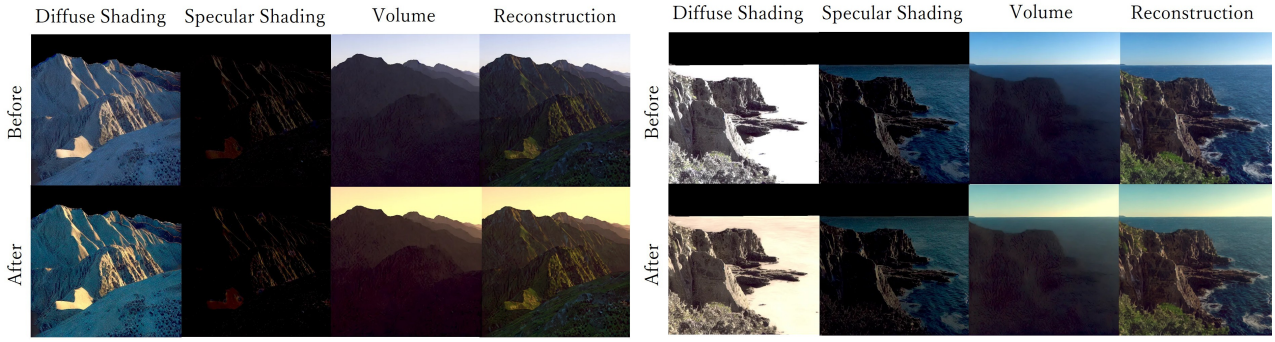


Figure 10: *Component-wise effects of atmospheric editing.*
Left: $\Delta\log(d_r, d_m, d_o) = (1.5, 1.5, 0)$*; Right:* $\Delta\log(d_r, d_m, d_o) = (1.5, 3.0, 0)$.

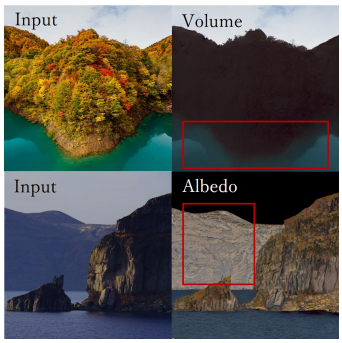


Figure 11: *Limitations of our image decomposition.*

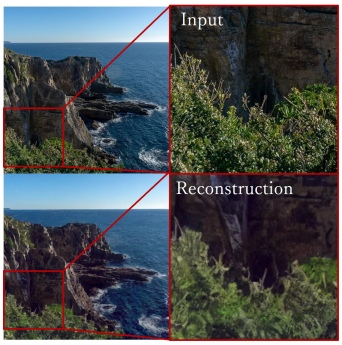


Figure 12: *Limitations of our reconstruction.*

5.7. Limitations

As shown in Fig. 11, our method may still suffer from leakage of water-region and fog effects into components where they should not be decomposed. One possible reason is that the model is trained only on our custom synthetic data, which may introduce bias in the data distribution. In addition, the reconstruction loss and reconstruction guidance introduced in our method do not guarantee perfect reconstruction of the input image. Therefore, as shown in Fig. 12, slight color differences may appear in the reconstructed image, and local structures may be lost, especially in vegetation regions. To address this problem, it would be desirable to design a network that predicts all components simultaneously so that the compositing equation in Eq. (4) is always satisfied. However, with the fine-tuning approach used in this study, the output channels of the U-Net cannot be easily expanded, as described in Section 4.3. Therefore, in this paper, we limit the correction to guidance during inference.

6. Conclusions

In this study, we proposed an intrinsic image decomposition method for large-scale natural terrain scenes that explicitly separates an input image into albedo, diffuse shading, specular shading, and an atmospheric volume component. By combining a synthetic dataset based on a physically based atmospheric scattering model with diffusion-based component estimation, reconstruction loss, and reconstruction guidance during inference, we showed that our method achieves better decomposition results than existing methods on synthetic evaluation datasets and plausible decomposition results on real-world outdoor images. Ablation studies further confirmed the effectiveness of each component of our method. In addition, by learning changes in atmospheric parameters in the component space, we demonstrated that our method can edit atmospheric appearance while preserving terrain details and reflection structures.

However, our method still has several limitations. First, water and sky masks for real images are obtained semi-automatically with manual correction, and automatic mask estimation remains future work. Second, quantitative evaluation on real-world images is limited because ground-truth intrinsic components and atmospheric parameters are difficult to obtain for real outdoor scenes. Therefore, establishing evaluation benchmarks for real-world images is important. In addition, due to the domain gap between synthetic and real images, water regions and fog effects may leak into other components. Furthermore, since each component is estimated independently, reconstruction loss and reconstruction guidance do not always guarantee perfect reconstruction of the input image. In future work, we aim to improve decomposition accuracy and reconstruction consistency by incorporating real-world data and improving the network architecture. We also plan to extend our system toward more flexible relighting by controlling the sun position in addition to atmospheric parameters.

Acknowledgments

This work was supported by JSPS KAKENHI JP23K28204, JP25K00154.

References

[BBS14] BELL S., BALA K., SNAVELY N.: Intrinsic images in the wild. *ACM Transactions on Graphics 33*, 4 (2014).

[BDL*21] BASLAMISLI A. S., DAS P., LE H.-A., KARAOGLU S., GEVERS T.: Shadingnet: Image intrinsics by fine-grained shading decomposition. *International Journal of Computer Vision 129*, 8 (2021). doi:10.1007/s11263-021-01477-5.

[BHE23] BROOKS T., HOLYNSKI A., EFROS A. A.: Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023).

[Ble] BLENDER KIT S.R.O.: Blenderkit: 3d models, materials, and other assets for blender. URL: https://www.blenderkit.com/.

[Ble26] BLENDER ONLINE COMMUNITY: *Blender: A 3D Modelling and Rendering Package*. Blender Foundation, 2026. Accessed: 2026-06-02. URL: https://www.blender.org/.

[CA23] CAREAGA C., AKSOY Y.: Intrinsic image decomposition via ordinal shading. *ACM Transactions on Graphics 43*, 1 (2023).

[CA24] CAREAGA C., AKSOY Y.: Colorful diffuse intrinsic image decomposition in the wild. *ACM Transactions on Graphics 43*, 6 (2024).

[DKG22] DAS P., KARAOGLU S., GEVERS T.: Pie-net: Photometric invariant edge guided network for intrinsic image decomposition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022).

[GJAF09] GROSSE R., JOHNSON M. K., ADELSON E. H., FREEMAN W. T.: Ground truth dataset and baseline evaluations for intrinsic image algorithms. In *Proceedings of the IEEE International Conference on Computer Vision* (2009).

[Hil20] HILLAIRE S.: A scalable and production ready sky and atmosphere rendering technique. *Computer Graphics Forum* (2020). doi:10.1111/cgf.14050.

[HZT*26] HAN X., ZHANG B., TANG X., LI X., WONKA P.: Lumix: Structured and coherent text-to-intrinsic generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2026).

[LDM*21] LE H.-A., DAS P., MENSINK T., KARAOGLU S., GEVERS T.: Eden: Multimodal synthetic dataset of enclosed garden scenes. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (2021).

[LGL*25] LIANG R., GOJCIC Z., LING H., MUNKBERG J., HASSELGREN J., LIN Z.-H., GAO J., KELLER A., VIJAYKUMAR N., FIDLER S., WANG Z.: Diffusionrenderer: Neural inverse and forward rendering with video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2025).

[LLLY24] LIN S., LIU B., LI J., YANG X.: Common diffusion noise schedules and sample steps are flawed. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (2024). doi:10.1109/WACV57701.2024.00532.

[LYS*21] LI Z., YU T.-W., SANG S., WANG S., SONG M., LIU Y., YEH Y.-Y., ZHU R., GUNDAVARAPU N., SHI J., BI S., YU H.-X., XU Z., SUNKAVALLI K., HAŠAN M., RAMAMOORTHI R., CHANDRAKER M.: Openrooms: An open framework for photorealistic indoor scene datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021). doi:10.1109/CVPR46437.2021.00711.

[RBL*22] ROMBACH R., BLATTMANN A., LORENZ D., ESSER P., OMMER B.: High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022).

[RRR*21] ROBERTS M., RAMAPURAM J., RANJAN A., KUMAR A., BAUTISTA M. A., PACZAN N., WEBB R., SUSSKIND J. M.: Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021).

[SP25] SARTOR S., PEERS P.: Teamwork: Collaborative diffusion with low-rank coordination and adaptation. In *ACM SIGGRAPH Asia Conference Proceedings* (2025).

[Tru] TRUE-VFX: True terrain 4. https://true-terrain-4.true-vfx.xyz/. Accessed: 2026-06-02.

[Wol] WOLF MARKET: Atmosphere. https://wolfmarket.gumroad.com/l/atmosphere. Accessed: 2026-06-02.

[WZZ*22] WANG T., ZHANG T., ZHANG B., OUYANG H., CHEN D., CHEN Q., WEN F.: Pretraining is all you need for image-to-image translation, 2022. arXiv:2205.12952.

[XPY*24] XI C., PENG S., YANG D., LIU Y., PAN B., LV C., ZHOU X.: Intrinsicanything: Learning diffusion priors for inverse rendering under unknown illumination, 2024. arXiv:2404.11593.

[ZDG*24] ZENG Z., DESCHAINTRE V., GEORGIEV I., HOLD-GEOFFROY Y., HU Y., LUAN F., YAN L.-Q., HAŠAN M.: Rgb↔x: Image decomposition and synthesis using material- and lighting-aware diffusion models. In *ACM SIGGRAPH 2024 Conference Papers* (2024). doi:10.1145/3641519.3657445.

[ZIE*18] ZHANG R., ISOLA P., EFROS A. A., SHECHTMAN E., WANG O.: The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018). doi:10.1109/CVPR.2018.00068.

[ZLH*22] ZHU J., LUAN F., HUO Y., LIN Z., ZHONG Z., XI D., WANG R., BAO H., ZHENG J., TANG R.: Learning-based inverse rendering of complex indoor scenes with differentiable monte carlo raytracing. In *SIGGRAPH Asia 2022 Conference Papers* (2022). doi:10.1145/3550469.3555407.