Intrinsic Omnidirectional Image Decomposition with Illumination Pre-Extraction

Rong-Kai Xu, Lei Zhang, *Member, IEEE,* and Fang-Lue Zhang, *Member, IEEE*

***Abstract*—Capturing an omnidirectional image with a 360-degree field of view entails capturing intricate spatial and lighting details of the scene. Consequently, existing intrinsic image decomposition methods face significant challenges when attempting to separate reflectance and shading components from a low dynamic range (LDR) omnidirectional images. To address this, our paper introduces a novel method specifically designed for the intrinsic decomposition of omnidirectional images. Leveraging the unique characteristics of the 360-degree scene representation, we employ a pre-extraction technique to isolate specific illumination information. Subsequently, we establish new constraints based on these extracted details and the inherent characteristics of omnidirectional images. These constraints limit the illumination intensity range and incorporate spherical-based illumination variation. By formulating and solving an objective function that accounts for these constraints, our method achieves a more accurate separation of reflectance and shading components. Comprehensive qualitative and quantitative evaluations demonstrate the superiority of our proposed method over state-of-the-art intrinsic decomposition methods.**

***Index Terms*—Intrinsic decomposition, omnidirectional image, reflectance, shading.**

I. Introduction

Omnidirectional images, also referred to as 360-degree panoramic images, offer a comprehensive representation of the entire scene by covering the full field-of-view. They have garnered considerable attention and are extensively utilized in diverse fields, including video surveillance, robotics, and virtual reality. The intrinsic decomposition of an omnidirectional image aims at extracting reflectance and shading components. This process effectively disentangles the intrinsic albedo and the interaction between light and geometry of the scene. The resulting decomposition can be leveraged for various image editing tasks, such as recoloring, retexturing, and relighting.

The problem of intrinsic omnidirectional image decomposition shares the same inherent ill-posedness as encountered in regular color images. Without prior knowledge of reflectance and shading, this problem remains unsolvable. However, the existing well-established assumptions regarding reflectance and shading, as employed by previous methods [1, 2, 3, 4], unfortunately, do not fully leverage the potential of omnidirectional properties. It has been shown that these assumptions are often insufficient in characterizing the complex spatial layout and lighting distribution present in 360-degree images, thus failing to fully exploit the special properties of omnidirectional images. Although a few deep learning-based methods have been proposed for intrinsic omnidirectional image decomposition [5, 6, 7, 8, 9], they heavily rely on a substantial amount of training data and encounter challenges in generalizing to real scenes.

An omnidirectional image captures a scene from all directions at a given viewpoint, enabling a holistic analysis of the relationships between all the pixels given relevant geometry information. This knowledge can be leveraged to determine the positions of possible light sources and their lighting effects in an omnidirectional image, addressing the challenge of estimating complex illumination distribution in previous intrinsic image decomposition methods. We propose to establish corresponding priors on the shading image by leveraging such lighting information. By providing better guidance for estimating the intensities in the shading image, the accuracy of the decomposed reflectance image can also be improved.

The primary contribution of this work is a novel method for the intrinsic decomposition of omnidirectional images. By exploiting the 360-degree geometric information of the scene provided by an omnidirectional image, we perform a pre-extraction of illumination information to estimate an initial shading distribution of the scene. Subsequently, we introduce new constraints on the illumination intensity and other shading terms for the shading image in our decomposition model. The experiments demonstrate the superiority of our proposed method over previous work on intrinsic omnidirectional image decomposition.

II. Related Work

According to the priors used in the intrinsic decomposition, current methods can be broadly classified into two categories: manual prior based methods and deep learning based methods. Because intrinsic decomposition is one of the sub-problems of inverse rendering, we also introduce the inverse rendering methods.

A. Methods based on manually designed priors

Due to the ill-posedness of the intrinsic image decomposition, most methods resort to manual priors about reflectance and shading in images [10]. The Retinex theory [11] is commonly used for setting the priors. It assumes that large and small image gradients correspond to the reflectance layer and shading layer respectively. Then, the two layers are separated by setting a threshold, which is heavily dependent on the selection of the threshold. Many variants of priors

Rong-Kai Xu and Lei Zhang are with the School of Computer Science, Beijing Institute of Technology, Beijing 100081, China.

Fang-Lue Zhang is with the School of Engineering and Computer Science, Victoria University of Wellington, Wellington 6012, New Zealand.

Manuscript received xxxx xx, 2023; revised xxxx xx, 2023.

[12, 13, 14, 15, 16] have also been proposed to guide the decomposition. Kimmel et al. [12] use a variational model that considers the shading layer to be smoothly varying in the space and the reflectance to be sparse. Then, it is solved as a quadratic programming problem by using Bayesian estimation, which can eventually generate better results on color images. Yue et al. [4] introduce a weighted L_1 norm constraint on neighboring pixels according to the color similarity, which can reduce the effect of illumination on reflectance by solving the problem in the HSV color space. Similarly, Bi et al. [17] propose to decompose image in the Lab color space and perform edge-preserving smoothing to remove some illumination effects in the decomposition process. Baslamisli and Gevers [18] improve the optimization framework based on a dense conditional random field [14] by introducing illumination invariant image descriptors. These priors are usually based on local constraints and do not take into account the relationship of shading and reflectance in the whole scene.

Besides, there are some non-local constraints applied to the intrinsic decomposition. Zhao et al. [13] introduce non-local constraints based on Retinex theory and texture analysis to identify points with similar properties at a distance. Garces et al. [19] cluster points with similar reflectance in the image to build a linear system describing the connections and relations between them. Likewise, Meka et al. [20, 21] perform K-means clustering by using histograms of images to obtain clustering results for reflectance, and introduce a prior on the consistency of non-local reflectance to obtain more accurate decomposition results.

Although the above methods are able to achieve good results on intrinsic decomposition of normal images, there is a lack of priors for omnidirectional images. Specifically, an omnidirectional field of view sees more illumination information than normal images. However, the above priors like shading smoothness only consider local spatial relationships and do not consider the full field of view, leading to a missed opportunity in utilizing this valuable information.

B. Deep learning based methods

Recent years have witnessed a dizzying rise of deep learning-based methods [22, 23, 9, 7, 8] used in intrinsic image decomposition. Narihira et al. [24] propose a CNN framework that can directly predict reflectance image as well as shading image from the input color image. Zhou et al. [25] modify the CNN framework to achieve joint estimation of reflectance, shading and normal. Some subsequent works like [26, 27, 8, 9] expand the training datasets and adapt the network structure and loss function to improve the decomposition quality. Shi et al. [28] propose a decomposition method for non-Lambertian materials, which constructs CNN networks by training a large amount of data with reflectance, shading and highlight information for generating better reflectance images.

Besides, Li et al. [5] propose a method to recover scene illumination, reflectance and geometry from a pair of stereo omnidirectional images, which enables the intrinsic decomposition of omnidirectional images. The illumination is obtained by reconstructing a near-field environmental light and the reflection is inferred by a deep learning model. Finally, the results are refined with the physical constraints between lighting and geometry. Although this method does the intrinsic decomposition of omnidirectional images, there is still insufficient intrinsic decomposition due to inaccurate estimation of reflectance by the neural network.

Generally, deep learning based methods are capable of learning potential priors based on a large amount of data for training, while they are usually incapable of generalizing. Although there are many datasets for intrinsic decomposition [1, 14, 29], there is a lack of relevant datasets that can be directly used for the intrinsic decomposition of omnidirectional images.

C. Inverse rendering

There are also some inverse rendering methods [6, 30, 31, 32, 33, 34, 35] that can recover the properties of the scene elements. Compared to intrinsic decomposition, the inverse rendering method implements a more specific separation of scene properties and allows to generate photo-realistic re-rendering results of the scene [36]. Li et al. [31] recover albedo, specular roughness and spatially-varying lighting of the scene using a deep inverse rendering framework. Zhu et al. [32] estimate these properties using dense vision transformer to perform greater photorealism results. Li et al. [6] extend the inverse rendering method to omnidirectional images and can handle more complex material. These methods recover more properties of the scene, and most of them use network to learn the priors of the scene. So the inverse rendering methods also have the shortcoming of learning based methods.

III. METHOD

A. Overview

Our method utilizes the holistic representation of the scene in a 360-degree image to pre-extract illumination information, which can effectively address the problem of inadequate decomposition of complex illumination encountered by existing methods. In Fig. 1, we provide an overview of our method.

Given an omnidirectional image as the input, our method first estimates the corresponding depth map using the method proposed by Wang et al. [37] and computes the normals based on the depth map. Subsequently, we extract bright regions that may contain highlights and light sources. The highlights and light sources are then filtered, and the shading image is obtained through shadow mapping and shading based on the positions of the extracted light sources. Next, we refine the estimation result to obtain the *pre-extraction* image. Furthermore, we formulate new priors associated with the shading image by imposing constraints based on both pre-extraction images and the inherent characteristics of omnidirectional images. These priors serve as the foundation for constructing a decomposition model, which can be formulated as a non-convex optimization problem. To solve this problem, we employ the alternating direction method of multipliers (ADMM) solver. Finally, we obtain the 360-degree reflectance and shading images from the input omnidirectional image.

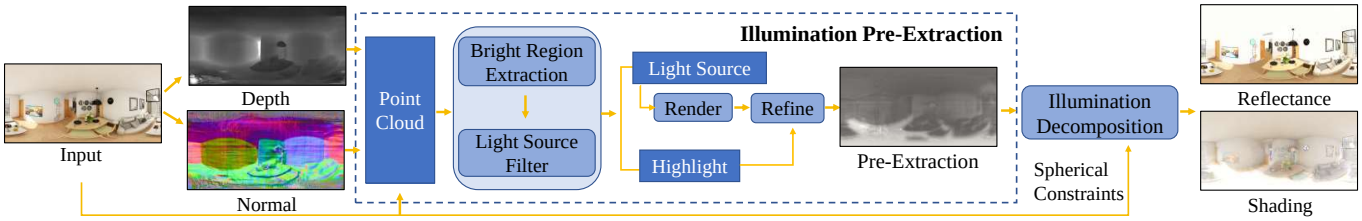


Fig. 1: Overview of our method. Our approach first estimates the depth and normal of the input LDR omnidirectional image. Then, we preform the illumination pre-extraction and decompose the input image into a reflectance image and a shading image. The constraints for our optimization framework are established based on the pre-extraction image and other inherent characteristics.

B. Decomposition model

Generally, the intrinsic image decomposition is defined as the following formulation:

$$\boldsymbol{I}(x) = \boldsymbol{R}(x) \times \boldsymbol{S}(x) \tag{1}$$

where $\boldsymbol{I}$ represents the input image, $\boldsymbol{R}$ and $\boldsymbol{S}$ represent the reflectance image and shading image respectively [36]. Here, $\times$ represents element-wise multiplication. For RGB color images, Eq. (1) is solved separately for each channel, so the channel index is omitted in the following expressions. To reduce the ambiguity in the decomposition process and attain a stable and rational outcome, it is essential to introduce additional constraints. These constraints are derived from observations of geometry, material properties, and illumination conditions.

The omnidirectional image provides a 360-degree representation of the scene, encompassing various aspects of illumination distribution such as interactions between the scene objects and light sources, diffuse reflections, specular reflections, and occlusions. By utilizing the 3D point cloud representation estimated from the 360-degree scene, the spatial relationships among all the visible scene points can be extracted, aiding in the inference of illumination information, such as the location of light sources, specular reflections, and shadows (see Fig. 2). This initial illumination information is employed to constrain the estimated illumination intensity of corresponding regions, aligning it with the actual intensity. This approach mitigates the problem of insufficient decomposition and enhances the results of both shading and reflectance image decomposition.

Overall, the decomposition model of Eq. (1) is constructed with the following two steps: the illumination pre-extraction that extracts the required initial illumination information from the omnidirectional image, and the objective function construction based on the pre-extracted illumination. Next, we provide a detailed explanation of these two steps, outlining their significance in the decomposition process.

Illumination Pre-Extraction. This pre-extraction process utilizes the point cloud generated from the depth and normal estimation to identify the locations of light sources and estimate the overall illumination distribution within the scene. Since the illumination is influenced by the relative positions of light sources, we initially extract bright regions that likely contain light sources and perform light source filtering to further refine the result. Subsequently, we estimate the illumination distribution based on the identified light sources. Through this process, we derive a reliable pre-extraction image that captures essential illumination characteristics and serves as a foundation for the subsequent decomposition.

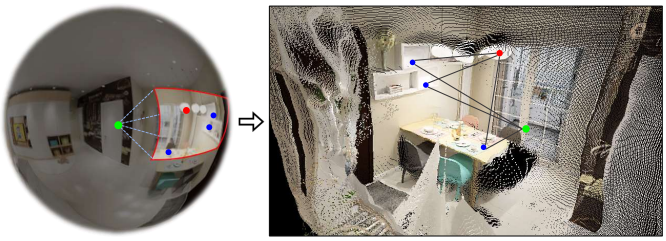


Fig. 2: Light source filtering. The viewpoint is indicated in green, the highlight is in blue, and the light source is in red. There are multiple reflected rays from different highlight regions that point to the light source, whereas there are usually few reflected rays from the light source to highlight region.

Bright Region Extraction. This stage aims to identify pixels with high intensity that are likely to represent highlights and light sources. To accomplish this, we employ the approach proposed by Shen et al. [38] to obtain the highlight-free result. Note that the extracted highlight areas contain not only the true highlight pixels, but also the light source pixels. Therefore, we get the bright region by subtracting the highlight-free image from the input image. These pixels in the top 2% of intensities are considered as the potential light sources, and used in the subsequent filtering process.

Light Source Filtering. Since directly distinguishing highlight and light source regions based on intensity alone is not feasible, a filtering step is employed at this stage. The fundamental principle behind this filtering approach is that if the light path from the viewer passes through a highlight region, it must also pass through the corresponding light source. Conversely, if the light path from the viewer passes through the light source, it does not necessarily pass through the highlight region. The reason for this difference is that highlights from the viewer's perspective are the result of specular reflection from the light source, and the light path is reversible. In Fig. 2, an example is provided to illustrate how reflected rays interact with highlight regions. This observation serves as the basis for determining whether a given region corresponds to a light source or a highlight region. By assessing the amount of light passing through the light path, it becomes possible to

differentiate between the two. The light sources and highlights can thus be filtered according to the number of reflected rays.

In practice, as the scene is represented by a sparse point cloud, we cannot directly use the ray-cast technique for intersection computation, so we choose the cone-cast technique as alternative. To simplify the calculation process, we only consider the case that the light is reflected once in the scene. For each pixel $\boldsymbol{p}$ extracted from the bright region, we record the direction from the viewpoint $\boldsymbol{v}$ to $\boldsymbol{p}$ as $\boldsymbol{V}_{\text{view}} = \boldsymbol{v} - \boldsymbol{p}$, and the direction of the specular reflection $\boldsymbol{V}_{\text{ref}}$ can be calculated based on the normal $\boldsymbol{N_p}$ as:

$$\boldsymbol{V}_{\text{ref}} = 2(\boldsymbol{N_p} \cdot \boldsymbol{V}_{\text{view}})\boldsymbol{N_p} - \boldsymbol{V}_{\text{view}} \tag{2}$$

Next, we determine the intersection point by intersecting the cone along the direction vector $\boldsymbol{V}_{\text{ref}}$ and count the number of these intersection points. It is important to note that the top angle of the cone defines the range of the reflection lobe. In order to achieve more accurate results, we have selected a top angle of $8°$ for the cone. By iterating through all the pixels in the image, we perform this cone intersection operation. The points that have the highest count of intersections are identified as potential light sources. This method enables us to locate prominent light sources such as windows, which significantly contribute to the overall illumination of the scene.

Then, we estimate the light color according to Gardner et al. [39]. Here we replace the light source position estimation stage with our own method mentioned above due to our LDR input. Concretely, we dilate the light source regions until the boundary intensity or gradient is less than the threshold , which is $0.8I_{max}$ for intensity and 0.05 for gradient in our implementation, to detect the full region of the light source and take the mean color of these regions as the light color.

Fig. 3: Examples of light source filtering. The light sources are shown in purple and overlaid on original images.

We present some visualized examples of the results obtained in the light source filtering stage in Fig. 3. These examples demonstrate the effectiveness of our approach in accurately identifying and localizing light sources within the scene. Besides, we show examples of light color estimation in Fig. 4. When evaluating the quality of our light color estimation, due to the lack of ground truth light source color, we create a color-biased image and compare the estimated light colors of the original image and its biased version. More specifically, we multiply a bias color C_b on the original image I_s to obtain its biased image I_b. Subsequently, we compute the color bias C_b^{est} between the estimated light source colors from I_s and I_b. Finally, we compare C_b^{est} and C_b to show that our method does not introduce additional errors when estimating the light color.

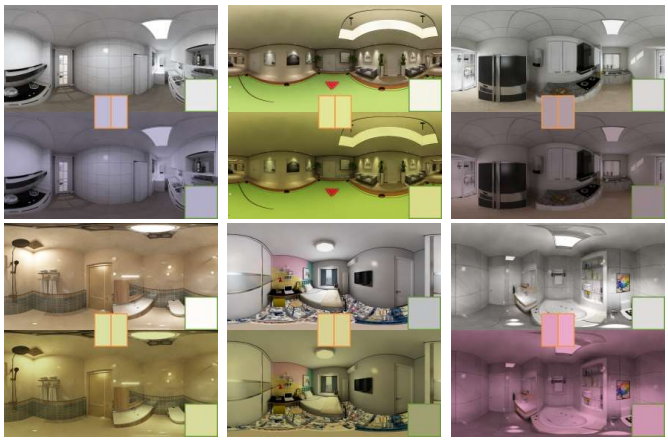


Fig. 4: Examples of light color estimation. Each case contains an origin image (top) and biased image (bottom), where the estimated light colors are shown in green boxes. The colors in the orange boxes indicate the real bias color of two images (left) and the bias of estimated color (right).

Illumination Estimation. With the light source information, we render the scene to get the illumination estimation [5]. For the pixel $\boldsymbol{p}$ with its normal $\boldsymbol{N_p}$, its shading is given by:

$$\boldsymbol{S_p} = \sum_{\boldsymbol{l} \in L} \frac{\max(\boldsymbol{V_l} \cdot \boldsymbol{N_p}, 0) + \max(\boldsymbol{V}_{\text{half}} \cdot \boldsymbol{N_p}, 0)^5}{\mathrm{D}(\boldsymbol{p}, \boldsymbol{l})^2} \tag{3}$$

In the equation, L represents the set of pixels corresponding to the identified light sources, and $\boldsymbol{V_l}$ denotes their respective light directions. The vector $\boldsymbol{V}_{\text{half}}$ is the half vector between the view direction and the light direction, while D represents the distance function used in the computation.

To incorporate the effects of shadows, we utilize a two-pass shadow mapping technique. This enables us to generate a pre-extraction image, denoted as $\boldsymbol{S}_{\text{est}}$, with $\boldsymbol{S_p}$ representing the pixel intensities.

However, it is important to acknowledge that the point cloud representation of the scene and the distortion in the high latitude area can introduce errors and artifacts in the results of the illumination estimation stage. Therefore, it is often necessary to refine the obtained illumination estimation results.

In particular, the pixels located in the high latitude area are susceptible to extreme distortions in their positions and normal directions, leading to inaccurate illumination estima-

tion. To address this issue, we replace these areas with the corresponding intensities from the input image as follows:

$$\boldsymbol{S}_{\text{est}} = \boldsymbol{\alpha}\boldsymbol{S}_{\text{est}} + (1-\boldsymbol{\alpha})\,|\boldsymbol{I}| \tag{4}$$

where $\boldsymbol{\alpha}$ is the weight corresponding to the latitude. Besides, the ceiling and floor are two areas with high latitude and we find that the floor usually receives more light than the ceiling. Hence, we increase the intensity of the floor and decrease the intensity of the ceiling by factor $\boldsymbol{\alpha}_{\text{layout}}$ as follows:

$$\boldsymbol{\alpha}_{\text{layout}}(\boldsymbol{p}) = \begin{cases} e^{2(\overline{\boldsymbol{I}}_{\text{ceiling}}-1)} & \text{if } \boldsymbol{p} \in \boldsymbol{M}_{\text{ceiling}} \\ e^{2\overline{\boldsymbol{I}}_{\text{floor}}} & \text{if } \boldsymbol{p} \in \boldsymbol{M}_{\text{floor}} \\ 1 & \text{otherwise} \end{cases} \tag{5}$$

where $\boldsymbol{M}_{\text{ceiling}}, \boldsymbol{M}_{\text{floor}}$ are the mask of ceiling and floor obtained by Xie et al. [40]. Note that we consider both the areas labeled as 'floor' and those with other labels, such as 'rug', due to their usually similar spatial positions. Here, $\overline{I}_{\text{ceiling}}$ and $\overline{I}_{\text{floor}}$ represent the mean intensity of the ceiling and floor areas respectively. The exponential function $e^{(\cdot)}$ maps the values into a valid range, i.e., smaller than 1 for the ceiling and larger than 1 for the floor. Here, we set the hyperparameter as 2 after experiments with different values. Finally, we normalize the intensity range of $\boldsymbol{S}_{\text{est}}$ to make sure there are no extreme values.

Fig. 5 shows some results of illumination pre-extraction. The pre-extraction image contains the illumination distribution of the scene, which can be used to establish new illumination intensity constraints.

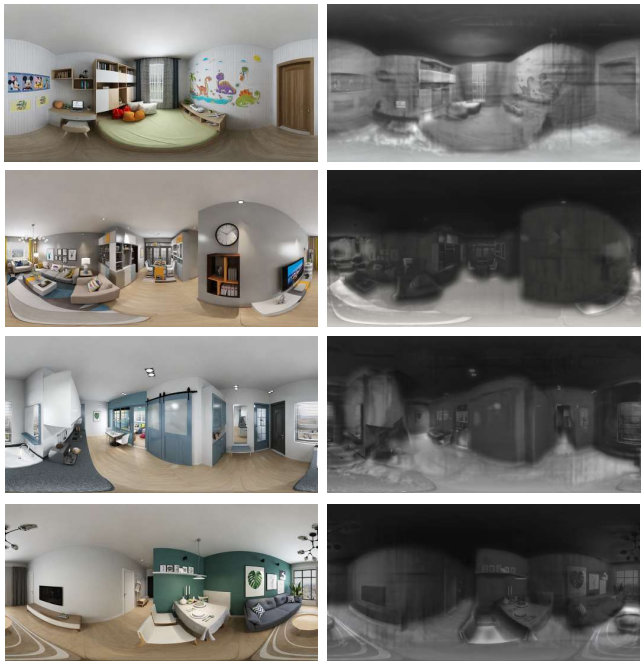


Fig. 5: Illumination pre-extraction. The left column displays the input images, and the right column displays the corresponding illumination pre-extraction image.

Objective function construction. After the illumination pre-extraction, we construct the decomposition model based on the pre-extraction results and some existing priors to solve Eq. (1). We formulate the intrinsic decomposition of omnidirectional image as an energy minimization problem with the following objective function:

$$E(\boldsymbol{R},\boldsymbol{S}) = E_{\text{data}}(\boldsymbol{R},\boldsymbol{S}) + E_{\text{ref}}(\boldsymbol{R},\boldsymbol{S}) + E_{\text{shading}}(\boldsymbol{R},\boldsymbol{S}) \tag{6}$$

which has three energy terms as the data term, the reflectance term and the shading term.

For the data term, we define the following weighted L_2 error metric to measure the reconstruction error.

$$E_{\text{data}}(\boldsymbol{R},\boldsymbol{S}) = \lambda \left\|(\boldsymbol{I} - \boldsymbol{R}\times\boldsymbol{S})\right\|_2^2 \tag{7}$$

where $\|\cdot\|_p$ denotes the p-norm operator.

And we follow the reflectance term defined in [41] that have the following expression:

$$E_{\text{ref}}(\boldsymbol{R},\boldsymbol{S}) = \lambda_{\text{r}} \left\|\nabla\boldsymbol{R}\right\|_0 \tag{8}$$

where ∇ is the gradient operator.

The shading term contains three sub-constraints, which describe the local constraint, non-local constraint, and illumination intensity constraint as follows:

$$E_{\text{shading}} = E_{\text{nonlocal}} + E_{\text{intensity}} \tag{9}$$

Non-Local Constraint. The non-local constraint use the pre-extraction image to achieve a stronger constraint on the shading image. The energy term is defined as follows:

$$E_{\text{nonlocal}}(\boldsymbol{R},\boldsymbol{S}) = \lambda_{\text{nls}} \sum_{(\boldsymbol{p},\boldsymbol{q})\in G} \boldsymbol{\omega}_{\text{nls}}(\boldsymbol{p},\boldsymbol{q}) \left(\boldsymbol{S}_{\boldsymbol{p}} - \boldsymbol{S}_{\boldsymbol{q}}\right)^2 \tag{10}$$

It is similar to the local constraint, but differs in the set G and weight $\boldsymbol{\omega}_{\text{nls}}$. The set G represents all point pairs with similar normal orientations $\boldsymbol{N}$ and spatial locations $\boldsymbol{P}$. To efficiently search for similar point pairs, we perform the search in a six-dimensional space comprising spatial positions and normal orientations. This choice helps prevent constraints between objects that are not spatially adjacent (see Fig. 6). Once the similar point pairs are identified, we calculate the weights for each pair based on the pre-extraction image. The weight computation follows the equation:

$$\boldsymbol{\omega}_{\text{nls}}(\boldsymbol{p},\boldsymbol{q}) = \begin{cases} \boldsymbol{\omega}_{\text{ls}}(\boldsymbol{p},\boldsymbol{q}) & \text{if } \left(\boldsymbol{S}_{\text{est},\boldsymbol{p}} - \boldsymbol{S}_{\text{est},\boldsymbol{q}}\right)^2 < \tau_{\text{nls}} \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

Here, $\boldsymbol{S}_{\text{est}}$ represents the pre-extraction image, and we set the threshold τ_{nls} to 0.05. This constraint allows us to avoid smoothing the shading of point pairs that possess similar positions and normals but exhibit inconsistent shading conditions (see Fig. 6). Specifically, when an object's surface is partially in shadow, this constraint ensures separate decomposition of shading inside and outside the shadow region. This separation improves the overall quality of the final result.

Illumination Intensity Constraint. The illumination intensity constraint is incorporated to ensure realistic results in the highlight and shadow regions of the shading image. In contrast to the method proposed by Bell et al. [14], where a constant is used to constrain the entire image, we utilize the pre-extraction image $\boldsymbol{S}_{\text{est}}$ to provide specific local constraints with respect to each pixel. Therefore, the following constraint is employed:

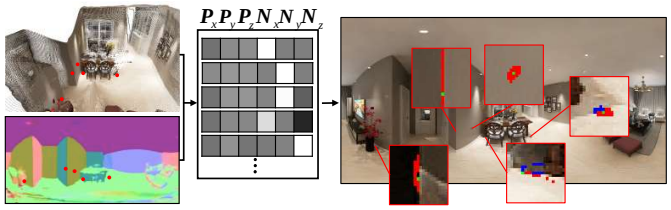


Fig. 6: Non-local constraint. The red pixels in the right image are pixels similar to the green ones, and the blue pixels are pixels excluded by the weight $\boldsymbol{\omega}_{\text{nls}}$.

$$E_{\text{intensity}}\left(\boldsymbol{R},\boldsymbol{S}\right)=\lambda_{\text{a}}\left\|\boldsymbol{S}-\boldsymbol{S}_{\text{est}}\right\|_2^2 \tag{12}$$

Since omnidirectional images capture a wider field of view compared to traditional images, the variation in shading becomes more comprehensive. The conventional approach of using a constant constraint to constrain the illumination intensity of the entire image fails to accurately capture the shading complexity present in omnidirectional images. On the other hand, the pre-extraction image provides a rough representation of the illumination intensity distribution within the scene. Consequently, it can be employed to constrain the illumination intensity of the shading image. This approach overcomes the limitation of the constant constraint, enabling a more accurate representation of the scene's illumination complexity.

C. Problem Solver

The objective function in Eq. (6) is non-convex due to the L_0 norm regularization, and contains two unknown variables. We use the alternating direction method of multipliers (ADMM) and a L_0 gradient minimization solution framework proposed by Xu et al. [42] to solve this optimization problem by introducing auxiliary variables. It divides the problem into several solvable convex optimization sub-problems. Also, for the convenience of formulating the problem, we rewrite the objective function in the following form:

$$\begin{aligned}\underset{\boldsymbol{R},\boldsymbol{S}}{\arg\min}&\lambda\left\|\left(\boldsymbol{I}-\boldsymbol{R}\times\boldsymbol{S}\right)\right\|_2^2+\lambda_{\text{r}}\left\|\nabla\boldsymbol{R}\right\|_0\\&+\lambda_{\text{nls}}\left\|\boldsymbol{M}_{\text{nls}}\boldsymbol{S}\right\|_2^2+\lambda_{\text{a}}\left\|\boldsymbol{S}-\boldsymbol{S}_{\text{est}}\right\|_2^2\end{aligned} \tag{13}$$

where $\boldsymbol{M}_{\text{nls}}$ are the matrix forms of the non-local constraint respectively. The number of rows of the matrix is the number of neighboring point pairs and the number of columns is the number of pixels. Each row of the matrix corresponds to a pair of points. If the points x and y are similar and the index of this point pair is p, then we have $\boldsymbol{M}\left(p,x\right)=\boldsymbol{\omega}_{\text{nls}}$ and $\boldsymbol{M}\left(p,y\right)=-\boldsymbol{\omega}_{\text{nls}}$, and the other elements in this row are set to 0.

We also introduce the auxiliary variable $\boldsymbol{G}=\nabla\boldsymbol{R}$ and an error variable $\boldsymbol{X}$, and rewrite Eq. (13) as the following form:

$$\begin{aligned}&\underset{\boldsymbol{R},\boldsymbol{S},\boldsymbol{G},\boldsymbol{X}}{\arg\min}\ \lambda\left\|\left(\boldsymbol{I}-\boldsymbol{R}\times\boldsymbol{S}\right)\right\|_2^2+\lambda_{\text{nls}}\left\|\boldsymbol{M}_{\text{nls}}\boldsymbol{S}\right\|_2^2\\&+\lambda_{\text{a}}\left\|\boldsymbol{S}-\boldsymbol{S}_{\text{est}}\right\|_2^2+\lambda_{\text{r}}\left\{\left\|\boldsymbol{G}\right\|_0+\mu\left\|\nabla\boldsymbol{R}-\boldsymbol{G}+\boldsymbol{X}\right\|_2^2\right\}\end{aligned} \tag{14}$$

According to the ADMM theory [43], Eq. (14) can be decomposed into processes about solving $\boldsymbol{R}$, $\boldsymbol{S}$, $\boldsymbol{G}$ and $\boldsymbol{X}$ individually, and the solution can be obtained by solving these variables iteratively towards the convergence.

IV. Experiments

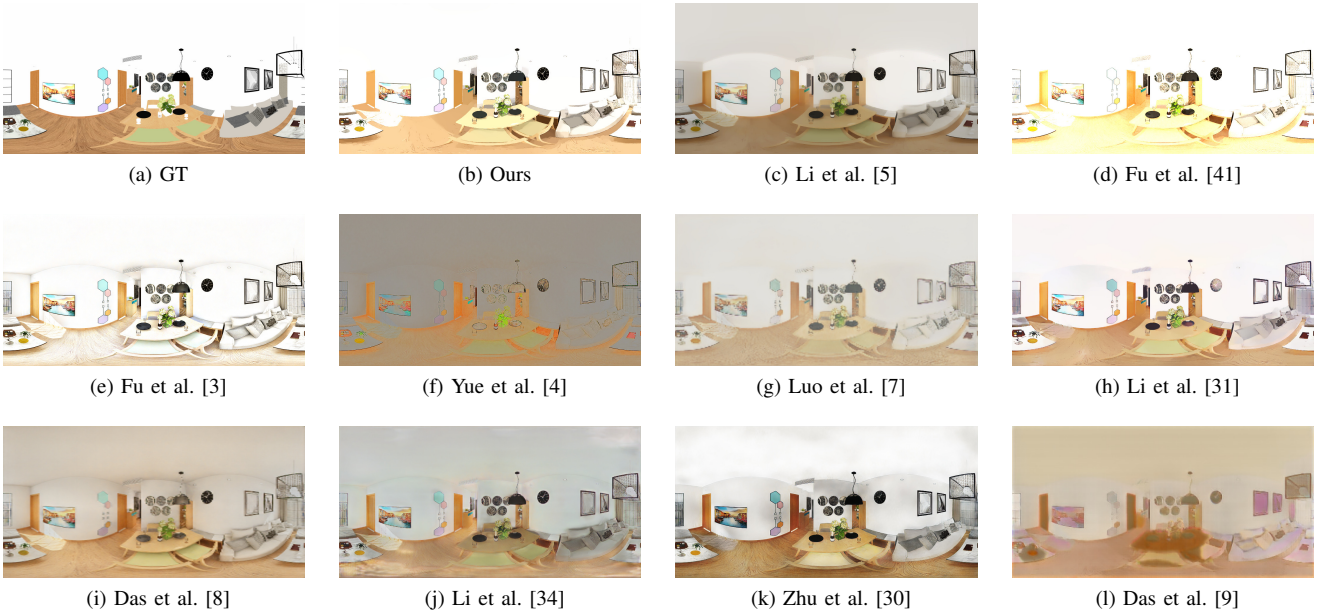
In this section, we evaluate the performance of our method in intrinsic omnidirectional image decomposition experiments. To assess its effectiveness, we compared our method against other intrinsic image decomposition and inverse rendering methods available in the literature. In our evaluation, we selected a range of techniques including manual priors-based methods such as those proposed by Fu et al. [3], Yue et al. [4] and Fu et al. [41], as well as learning-based methods such as those introduced in Li et al. [5], Das et al. [8], Li et al. [34], Luo et al. [7], Li et al. [31], Zhu et al. [30], Li et al. [6] and Das et al. [9]. Since some methods are not explicitly crafted for omnidirectional images, we devised two approaches to ensure a fair comparison. In the first approach, we directly input the omnidirectional image into these methods and extract the decomposition results. The second approach involves projecting the omnidirectional image onto a cubemap, treating each face as a traditional perspective image. Subsequently, we conduct intrinsic decomposition on each face of the cubemap. After this process, we project the results back onto the omnidirectional image to facilitate a comprehensive comparison with our method and other omnidirectional decomposition methods.

There are several parameters related to solving the objective function in our proposed method, and we set $\lambda=0.3$, $\lambda_{\text{r}}=0.005$, $\lambda_{\text{a}}=0.02$, $\lambda_{\text{nls}}=0.005$ for the experiments in this paper. We implemented our method by programming with Matlab on a computer with an Intel i7-11700 2.5GHz CPU and 32GB RAM. Our method takes about 25.6s to decompose an omnidirectional image with the resolution of 256×512, 102.2s for 512×1024 and 422.7s for 1024×2048 without specific acceleration. Since most parts of our algorithm can be implemented in parallel, including illumination extraction and model solving, the performance of our method can be significantly improved by GPU acceleration. Next, we elaborate the details of the experiments.

A. Decomposition Results

We use synthetic data [44] and real data [5, 45], and preform both qualitative and quantitative analysis on the results obtained by our method and state-of-the-art methods. Fig. 7 and Fig. 8 show some examples of the synthetic data. Fig. 9 and Fig. 10 show the results on the real-world data. Note that all the results shown here are obtained by the aforementioned approach that does not project deocomposition results to cubemap. Please referred to the supplementary material for more results of both the two approaches (*supplementary.pdf*). Generally, our method can achieve better results than other methods.

Qualitative analysis. As illustrated in Fig. 7 and Fig. 8, the decomposition method proposed by Yue et al. [4] fails to accurately estimate the reflectance of omnidirectional

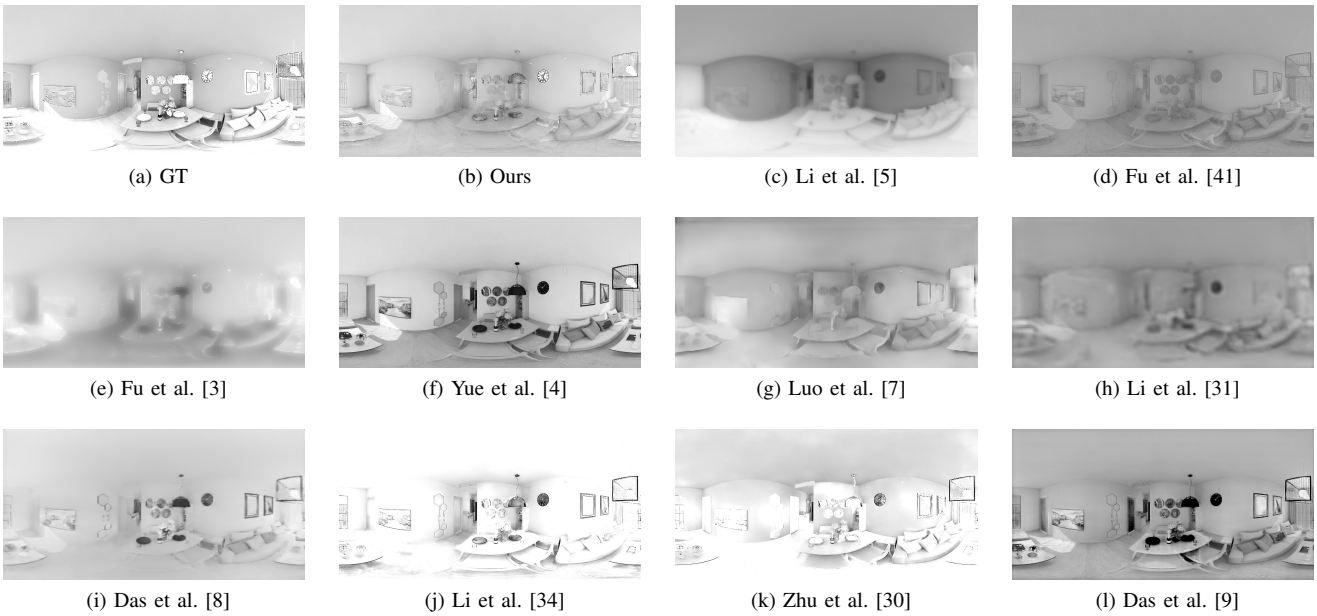


(a) GT (b) Ours (c) Li et al. [5] (d) Fu et al. [41]

(e) Fu et al. [3] (f) Yue et al. [4] (g) Luo et al. [7] (h) Li et al. [31]

(i) Das et al. [8] (j) Li et al. [34] (k) Zhu et al. [30] (l) Das et al. [9]

Fig. 7: Reflectance results of examples in the Structure3D dataset [44].



(a) GT (b) Ours (c) Li et al. [5] (d) Fu et al. [41]

(e) Fu et al. [3] (f) Yue et al. [4] (g) Luo et al. [7] (h) Li et al. [31]

(i) Das et al. [8] (j) Li et al. [34] (k) Zhu et al. [30] (l) Das et al. [9]

Fig. 8: Shading results of examples in the Structure3D dataset [44]. For better visualization, we show all results in grayscale.

images. It does not yield the correct reflectance values for the input image. The method developed by Das et al. [8] shows only minimal improvements and remains close to the input image without achieving significant decomposition. Luo et al. [7] obtain reasonable results of single object with constant reflectance, but shows the inconsistent results between different objects. For example, the wall is darker than ground truth while the wooden floor is brighter. In the case of other intrinsic decomposition methods such as [41, 3], they exhibit partial success in recovering certain parts of the scene, such as the ceiling and walls. However, these methods struggle to estimate the reflectance of the floor accurately. This limitation arises because they rely solely on local information, neglecting the crucial omnidirectional information that provides a holistic description of the entire scene. While inverse rendering methods can also recover reflectance images, some of them [34, 30], do not perform well when applied to omnidirectional images. These methods generate reflectance estimates that are inconsistent and contain artifacts, such as variations in reflectance on the wall. Although the methods proposed by Li et al. [31, 5] produce better results compared to others, they still exhibit inconsistencies in reflectance variation.

Our approach is specifically designed to address the challenges posed by omnidirectional images, with the primary

(a) Input (b) Ours (c) Li et al. [5] (d) Li et al. [6] (e) Li et al. [31]

(f) Fu et al. [41] (g) Luo et al. [7] (h) Das et al. [8] (i) Zhu et al. [30] (j) Li et al. [34]

Fig. 9: Results of examples in the dataset [5].

(a) Input (b) Ours (c) Li et al. [31] (d) Fu et al. [41]

(e) Li et al. [34] (f) Luo et al. [7] (g) Das et al. [8] (h) Zhu et al. [30]

Fig. 10: Results of examples in the Stanford2D3D dataset [45].

goal of accurately estimating scene illumination and mitigating decomposition ambiguity. This leads to more precise and reliable outcomes. This improvement is evident in Fig. 7(b) and Fig. 8(b), where our method surpasses other techniques in producing superior reflectance and shading images. Notably, our approach achieves excellent alignment between the estimated reflectance of the ceiling and walls and the ground truth, showcasing the accuracy of our method. Furthermore, our method consistently delivers reliable results for various objects within the scene, including doors and floors, benefiting from the utilization of global scene information.

Fig. 9 shows the results on some real data [5]. Here we also make the comparison with the method of Li et al. [6] [1], which is proposed for panoramic inverse rendering. Our method can also have better performance than other methods on the real data.

As shown in Fig. 9(b), our method can remove the illumination effects more accurately than other methods [5, 41, 8], such as effects on the green wall, the floor and the sofa. Additionally, our method achieves reflectance sparsity, ensuring that the reflectance of the same objects remains as consistent as possible. Some results obtained by other methods [31, 34, 30, 6] have large reflectance variation than ours. Specifically, due to the spherical constraints used in the local constraint, the reflectance obtained by our method tends to be more consistent in high latitude areas, which is more accurate since these points occupy very small area in the scene.

In Fig. 10, we present some results obtained on the Stanford2D3D dataset [45], which exhibits more complex illumination variations compared to other datasets. It is observed that most existing methods [30, 34, 31, 41], struggle to achieve reliable decomposition in such challenging scenarios. For instance, they fail to capture the darkened ceiling resulting from insufficient illumination. In contrast, our method produces superior results due to the incorporation of layout information during the illumination pre-extraction stage. This additional information aids in accurately decomposing high-latitude areas like the ceiling and floor.

The qualitative analysis above clearly demonstrates the significant advantages of our method in performing intrinsic decomposition on omnidirectional images, using both synthetic and real data. To further illustrate these advantages, we will now proceed with a quantitative analysis.

Quantitative analysis. For quantitative analysis, we utilize synthetic omnidirectional images from the Structure3D dataset [44]. This dataset comprises indoor scenes with complex lighting and geometry, providing a suitable testbed for evaluating our algorithm. However, certain scenes in this dataset contain transparent objects, where the reflectance is considered white. Standard intrinsic decomposition algorithms do not account for the effect of transparent objects and may present the

[1]Because the code is not available, we only compare with it on the data used in the paper, and the results are directly obtained from the paper [6].

TABLE I: Quantitative evaluation of the reflectance (R) and shading (S) images by different methods in cubemap.

Method	sMSE↓			sSMSE↓			sLMSE↓			SSIM↑		
	R	S	Mean	R	S	Mean	R	S	Mean	R	S	Mean
Yue et al. [4]	0.1409	0.0925	0.1167	0.0690	0.0721	0.0705	0.0475	0.0663	0.0569	0.6133	0.5727	0.5930
Fu et al. [3]	0.0460	0.0468	0.0464	0.0366	0.0406	0.0386	0.0279	0.0380	0.0330	0.6368	0.5575	0.5971
Das et al. [8]	0.0599	0.0639	0.0619	0.0421	0.0464	0.0443	0.0306	0.0407	0.0356	0.6443	0.5587	0.6015
Li et al. [5]	-	-	-	-	-	-	-	-	-	-	-	-
Fu et al. [41]	0.0592	0.0890	0.0741	0.0367	0.0308	0.0338	0.0269	0.0284	0.0276	0.6810	0.5523	0.6166
Luo et al. [7]	0.0447	0.0415	0.0431	0.0373	0.0327	0.0350	0.0282	0.0284	0.0283	0.7169	0.5746	0.6457
Zhu et al. [30]	0.2329	0.1164	0.1746	0.1229	0.0380	0.0804	0.0783	0.0317	0.0550	0.5887	0.5727	0.5807
Li et al. [31]	0.0701	0.0710	0.0706	0.0484	0.0342	0.0413	0.0378	0.0313	0.0345	0.4125	0.3985	0.4055
Li et al. [34]	0.0723	0.0546	0.0634	0.0384	0.0359	0.0371	0.0274	0.0326	0.0300	0.6772	0.6479	0.6625
Das et al. [9]	0.0788	0.1280	0.1034	0.0612	0.0982	0.0797	0.0467	0.1044	0.0756	0.4363	0.5480	0.4922
Ours	-	-	-	-	-	-	-	-	-	-	-	-

TABLE II: Quantitative evaluation of the reflectance (R) and shading (S) images by different methods in omnidirectional image.

Method	sMSE↓			sSMSE↓			sLMSE↓			SSIM↑		
	R	S	Mean	R	S	Mean	R	S	Mean	R	S	Mean
Yue et al. [4]	0.1388	0.1148	0.1268	0.0667	0.0638	0.0653	0.0455	0.0617	0.0536	0.6255	0.5688	0.5972
Fu et al. [3]	0.0534	0.0545	0.0539	0.0404	0.0427	0.0416	0.0294	0.0365	0.0330	0.6973	0.5533	0.6253
Das et al. [8]	0.0892	0.0549	0.0721	0.0438	0.0418	0.0428	0.0326	0.0392	0.0359	0.6383	0.5533	0.5958
Li et al. [5]	0.0564	0.0753	0.0658	0.0332	0.0575	0.0454	**0.0213**	0.0333	0.0273	0.7308	0.5489	0.6398
Fu et al. [41]	0.0562	0.0713	0.0638	0.0347	0.0288	0.0318	0.0263	0.0276	0.0270	0.6842	0.5648	0.6245
Luo et al. [7]	0.0445	0.0478	0.0461	0.0378	0.0308	0.0343	0.0313	0.0307	0.0310	0.6912	0.5495	0.6204
Zhu et al. [30]	0.1062	0.0745	0.0904	0.0548	0.0292	0.0420	0.0371	**0.0247**	0.0309	0.6969	0.6465	0.6717
Li et al. [31]	**0.0328**	0.0705	0.0517	**0.0284**	0.0309	0.0297	0.0222	0.0292	0.0257	0.7163	0.6072	0.6618
Li et al. [34]	0.0791	0.0560	0.0675	0.0313	0.0312	0.0312	0.0243	0.0298	0.0271	0.6995	0.6691	0.6843
Das et al. [9]	0.0711	0.1023	0.0867	0.0511	0.0649	0.0580	0.0414	0.0771	0.0592	0.4246	0.5454	0.4850
Ours	0.0367	**0.0312**	**0.0339**	0.0285	**0.0261**	**0.0273**	0.0241	0.0249	**0.0245**	**0.7552**	**0.7035**	**0.7293**

reflectance of objects behind them. This can introduce errors during quantitative analysis for all the methods, particularly when transparent objects occupy a significant portion of the image. To ensure accurate evaluation, we exclude scenes with a substantial number of transparent objects from our experiments. In total, we select 483 images for quantitative analysis.

To evaluate our method, we employ four commonly used metrics in intrinsic decomposition methods, including mean square error (MSE), scale-invariant mean square error (SMSE), local mean square error (LMSE), and the structure similarity index measure (SSIM). Since omnidirectional images are projected from a sphere, each pixel in the omnidirectional image may contribute differently to the final metric. Therefore, we incorporate solid angle-based weights into the three MSE metrics to obtain spherical MSE metrics denoted as sMSE, sSMSE, and sLMSE, following Weber et al. [46]'s method. Additionally, we calculate errors separately for reflectance and shading images.

Table I and Table II present the error statistics for different methods. The best results are indicated in bold, while results close to the best are underlined. In terms of mean error metrics, our method outperforms both manual prior-based methods and learning-based methods across all four metrics. Some methods [5, 31, 34], primarily focus on optimizing reflectance loss, which can result in better reflectance images but poorer shading images.

B. Ablation Study

To evaluate the effectiveness of the proposed constraints in our work, we conducted an ablation study to compare decomposition results with and without these constraints.

In the illumination pre-extraction stage, we compare the results obtained with and without the pre-extraction image. Specifically, the output of the illumination pre-extraction stage is represented as $\boldsymbol{S}_{\text{est}}$ in Eq. (12). The method without pre-extraction is implemented by using the illumination intensity constraint introduced in [41], where $\boldsymbol{S}_{\text{est}}$ in Eq. (12) is replaced with a constant value of $\boldsymbol{S}_{\text{est}} = 0.5$. Additionally, we include the method with $\boldsymbol{S}_{\text{est}} = |\boldsymbol{I}|$ for comparison, as the intensity image of the input often serves as an initial estimation of the shading image. Furthermore, to assess the effectiveness of the proposed refinement in the illumination pre-extraction stage, we compare results with and without refinement.

Regarding the energy terms, we compare methods with and without each term individually to demonstrate the necessity and validity of the energy terms introduced in our paper. Specifically, we set $\lambda_{\text{nls}} = 0$ to obtain results without the non-local shading constraints of E_{nonlocal}. Here, $\lambda_{\text{a}} = 0$ represents results without the illumination intensity term ($E_{\text{intensity}}$).

As presented in Table III, the illumination pre-extraction stage significantly improves the decomposition results. The

TABLE III: Ablation study of the reflectance (R) and shading (S) images.

Method	sMSE↓			sSMSE↓			sLMSE↓			SSIM↑		
	R	S	Mean	R	S	Mean	R	S	Mean	R	S	Mean
$\boldsymbol{S}_{\text{est}} = 0.5$	0.0509	0.0371	0.0440	0.0330	0.0269	0.0300	0.0256	0.0253	0.0254	0.6912	0.6244	0.6578
$\boldsymbol{S}_{\text{est}} = \lvert\boldsymbol{I}\rvert$	0.0453	0.0371	0.0412	0.0337	0.0313	0.0325	0.0252	0.0274	0.0263	0.6566	0.5908	0.6237
w/o $\boldsymbol{S}_{\text{est}}$ refinement	0.0417	0.0348	0.0383	0.0315	0.0278	0.0297	0.0246	0.0257	0.0251	0.6901	0.6490	0.6696
w/o E_{nonlocal}	0.0382	0.0326	0.0354	0.0291	0.0270	0.0281	0.0243	0.0259	0.0251	0.7453	0.7022	0.7237
w/o $E_{\text{intensity}}$	0.0896	0.0806	0.0851	0.0400	0.0298	0.0349	0.0289	0.0266	0.0277	0.6705	0.5840	0.6273
Ours	**0.0367**	**0.0312**	**0.0339**	**0.0285**	**0.0261**	**0.0273**	**0.0241**	**0.0249**	**0.0245**	**0.7552**	**0.7035**	**0.7293**

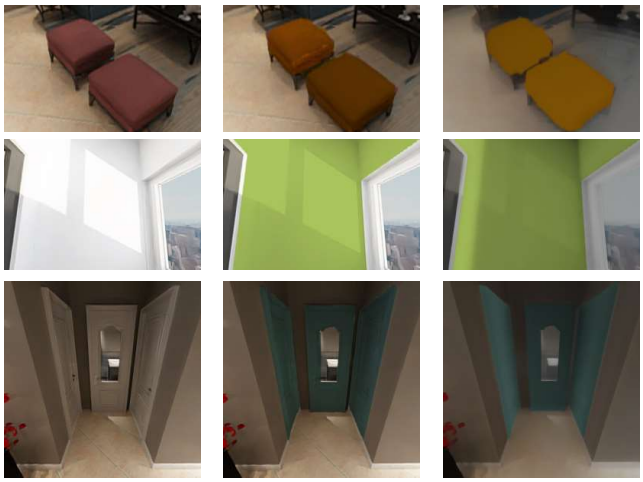
introduced constraints are demonstrated to be necessary and effective in achieving accurate decomposition and enhancing the quality of decomposition.

C. Applications

We show some image editing applications based on our intrinsic omnidirectional image decomposition results, including recoloring, retexturing and relighting. Also we show the results of Li et al. [5], which works for intrinsic omnidirectional image decomposition method. Our method consistently produces more realistic results, whether it involves modifying the reflectance or shading image.

Recoloring. Our method enables realistic recoloring of different objects within an image. Given an input omnidirectional image, we first perform intrinsic decomposition to obtain the reflectance image and shading image. We then modify the value of a selected region in the reflectance image to change the color of the objects within that region. Finally, we generate the recolored image by multiplying the modified reflectance image with the shading image.

In Fig. 11, we present an example of recoloring using our decomposition results, which outperform existing intrinsic decomposition methods. Notably, in the second row, our method successfully preserves the light effects that are missing in the method proposed by Li et al. [5]. This observation indicates that our method accurately separates the shading properties from the input image, resulting in a more faithful preservation of light effects during the recoloring process.



(a) Input (b) Ours (c) Li et al. [5]

Fig. 11: The recoloring results obtained by using the decomposition results of our method and Li et al. [5].

Retexturing. Retexturing and recoloring share a similarity in that they both involve modifying the reflectance image during image editing. However, there is a distinction between the two processes. In retexturing, the ideal reflectance image does not contain any shading information. Therefore, the texture used for retexturing does not need to consider shading effects. The result of retexturing can be obtained by leveraging the shading image to achieve a natural replacement of the texture. In Fig. 12, we present an example of retexturing based on our decomposition results. Our method excels in achieving more realistic texture editing outcomes.

(a) Input (b) Ours (c) Li et al. [5]

Fig. 12: The retexturing results obtained by using the decomposition results of our method and Li et al. [5]. Some decorative paintings are added on the wall.

Relighting. Our method offers the ability to incorporate new light sources into a scene. Leveraging the point cloud representation of the scene geometry obtained through depth estimation, we can calculate the shading image corresponding to the new light source. By combining this shading image with the original shading image obtained from intrinsic decomposition, we generate the result with the added light source. Fig.

13 showcases the relighting results achieved using our method in various indoor scenes.

(a) Input (b) Ours (c) Li et al. [5]

Fig. 13: The relighting results by using the decomposition results of our method and Li et al. [5]. Different light sources are added into the scenes, such as the point and spot lights.

D. Limitation and discussion

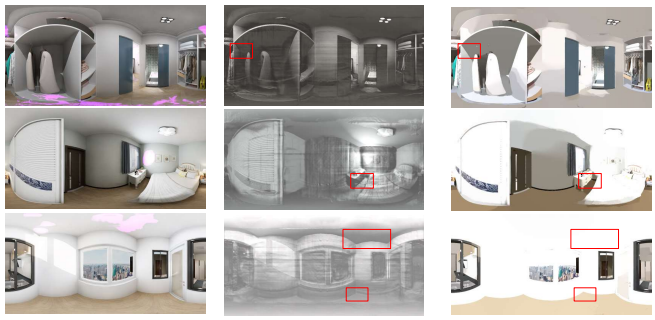
The experiments above show that our method outperforms other methods on LDR omnidirectional images decomposition. Since our method doesn't depend on the network, it can also work for other situations like HDR omnidirectional images. Fig. 14 shows the examples on HDR dataset. While our method may not achieve the precise reflectance, it does succeed in capturing reflectance with certain shading effects removed, while also preserving the boundaries between different objects, such as windows. This indicates that our method is able to generalise to HDR inputs. While Li et al. [31]'s method is completely unable to deal with HDR input and does not give reasonable results.

(a) Input (b) Ours (c) Li et al. [31]

Fig. 14: Results on the Laval HDR Indoor dataset [47].

While our method demonstrates good performance on many omnidirectional datasets, there are still several aspects of this complex ill-posed problem that remain unexplored. For example, there are still some shadow and highlight that are not fully removed. This is a challenge prevalent in many intrinsic decomposition methods. Despite our method's pre-extraction of shading, it struggles to comprehensively depict the illumination distribution in scenes, particularly in regions with limited geometry information. Additionally, specific reflectance details, such as wood grain, are inadvertently eliminated. This is a consequence of the sparse reflectance prior, where reflectance is treated as constant within small patches. As a result, the decomposition results tend to be piece-wise constant, leading to the removal of subtle details in the process. The problem also exists for other methods that use this prior [3, 4, 41]. Besides, our method relies on the illumination pre-extraction step based on the estimated depth and normal maps. However, it is important to note that the accuracy of these estimated depth and normal values is not always guaranteed. This might result in incorrect pre-extracted illumination, particularly in scenes that primarily consist of small light sources such as spotlights and lamps.

We present some failure cases in Fig. 15, which highlight instances where the placement of the light source is incorrect, leading to inaccurate illumination (indicated by the red box in the pre-extraction stage). Consequently, certain illumination effects persist in the reflectance image. Nonetheless, it is worth mentioning that our method incorporates shading and reflectance constraints, which help mitigate the impact of these inaccuracies in the pre-extraction stage. This can be observed in the third row of the results, where our method partially rectifies the effects of the inaccurate pre-extraction results.



(a) Input&Light Source (b) Pre-extraction (c) Reflectance

Fig. 15: The incorrect light source position (highlighted in purple) causes inaccurate illumination pre-extraction (red boxes).

V. Conclusion

We have presented a novel intrinsic decomposition method specifically designed for omnidirectional images. By leveraging the unique characteristic of omnidirectional images, which provide a 360-degree representation of the scene, we introduced an illumination pre-extraction step to address the limitations of existing priors in accurately describing scene illumination. This pre-extraction step significantly enhances the quality of the decomposition results. The experimental evaluations have shown the effectiveness of our method, showcasing its potential for various image editing applications.

In terms of future work, we intend to delve into more accurate and robust methods for extracting semantic information about the scene, leveraging deep learning techniques. This includes exploring techniques to extract information such as light and object categories, as well as occlusion relations.

Additionally, extending our method to intrinsic decomposition of omnidirectional videos holds promising potential.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their helpful comments. This work was supported by the National Natural Science Foundation of China under Grant 62132012. This work was also partly supported by the Marsden Fund Council managed by the Royal Society of New Zealand (No. MFP-20-VUW-180).

REFERENCES

[1] R. Grosse, M. K. Johnson, E. H. Adelson, and W. T. Freeman, "Ground truth dataset and baseline evaluations for intrinsic image algorithms," in *Proceedings of ICCV*, 2009, pp. 2335–2342.

[2] J. Shen, X. Yang, Y. Jia, and X. Li, "Intrinsic images using optimization," in *Proceedings of CVPR*, 2011, pp. 3481–3487.

[3] X. Fu, D. Zeng, Y. Huang, X.-P. Zhang, and X. Ding, "A weighted variational model for simultaneous reflectance and illumination estimation," in *Proceedings of CVPR*, 2016, pp. 2782–2790.

[4] H. Yue, J. Yang, X. Sun, F. Wu, and C. Hou, "Contrast enhancement based on intrinsic image decomposition," *IEEE Transactions on Image Processing*, vol. 26, no. 8, pp. 3981–3994, 2017.

[5] J. Li, H. Li, and Y. Matsushita, "Lighting, reflectance and geometry estimation from 360 panoramic stereo," in *Proceedings of CVPR*, 2021, pp. 10 586–10 595.

[6] Z. Li, L. Wang, X. Huang, C. Pan, and J. Yang, "PhyIR: physics-based inverse rendering for panoramic indoor images," in *Proceedings of CVPR*, 2022, pp. 12 713–12 723.

[7] J. Luo, Z. Huang, Y. Li, X. Zhou, G. Zhang, and H. Bao, "NIID-Net: Adapting surface normal knowledge for intrinsic image decomposition in indoor scenes," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 12, pp. 3434–3445, 2020.

[8] P. Das, S. Karaoglu, and T. Gevers, "PIE-Net: Photometric invariant edge guided network for intrinsic image decomposition," in *Proceedings of CVPR*, 2022, pp. 19 790–19 799.

[9] P. Das, S. Karaoğlu, A. Gijsenij, and T. Gevers, "SIGNet: Intrinsic image decomposition by a semantic and invariant gradient driven network for indoor scenes," in *Proceedings of ECCV*, 2023, pp. 605–620.

[10] N. Bonneel, B. Kovacs, S. Paris, and K. Bala, "Intrinsic decompositions for image editing," *Computer Graphics Forum*, vol. 36, no. 2, pp. 593–609, 2017.

[11] E. H. Land and J. J. McCann, "Lightness and Retinex theory," *Journal of the Optical Society of America*, vol. 61, no. 1, pp. 1–11, 1971.

[12] R. Kimmel, M. Elad, D. Shaked, R. Keshet, and I. Sobel, "A variational framework for Retinex," *International Journal of Computer Vision*, vol. 52, no. 1, pp. 7–23, 2003.

[13] Q. Zhao, P. Tan, Q. Dai, L. Shen, E. Wu, and S. Lin, "A closed-form solution to Retinex with nonlocal texture constraints," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 7, pp. 1437–1444, 2012.

[14] S. Bell, K. Bala, and N. Snavely, "Intrinsic images in the wild," *ACM Transactions on Graphics*, vol. 33, no. 4, pp. 1–12, 2014.

[15] W. Ma and S. Osher, "A TV bregman iterative model of retinex theory," *Inverse Problems & Imaging*, vol. 6, no. 4, p. 697, 2012.

[16] W. Ma, J.-M. Morel, S. Osher, and A. Chien, "An L1-based variational model for retinex theory and its application to medical images," in *Proceedings of CVPR*, 2011, pp. 153–160.

[17] S. Bi, X. Han, and Y. Yu, "An L1 image transform for edge-preserving smoothing and scene-level intrinsic decomposition," *ACM Transactions on Graphics*, vol. 34, no. 4, pp. 1–12, 2015.

[18] A. S. Baslamisli and T. Gevers, "Invariant descriptors for intrinsic reflectance optimization," *Journal of the Optical Society of America*, vol. 38, no. 6, pp. 887–896, 2021.

[19] E. Garces, A. Munoz, J. Lopez-Moreno, and D. Gutierrez, "Intrinsic images by clustering," *Computer Graphics Forum*, vol. 31, no. 4, pp. 1415–1424, 2012.

[20] A. Meka, M. Zollhöfer, C. Richardt, and C. Theobalt, "Live intrinsic video," *ACM Transactions on Graphics*, vol. 35, no. 4, pp. 1–14, 2016.

[21] A. Meka, M. Shafiei, M. Zollhöfer, C. Richardt, and C. Theobalt, "Real-time global illumination decomposition of videos," *ACM Transactions on Graphics*, vol. 40, no. 3, pp. 1–16, 2021.

[22] T. Nestmeyer and P. V. Gehler, "Reflectance adaptive filtering improves intrinsic image estimation," in *Proceedings of CVPR*, 2017, pp. 6789–6798.

[23] Y. Qian, M. Shi, J.-K. Kamarainen, and J. Matas, "Fast fourier intrinsic network," in *Proceedings of WACV*, 2021, pp. 3169–3178.

[24] T. Narihira, M. Maire, and S. X. Yu, "Direct intrinsics: Learning albedo-shading decomposition by convolutional regression," in *Proceedings of ICCV*, 2015, pp. 2992–2992.

[25] H. Zhou, X. Yu, and D. W. Jacobs, "Glosh: Global-local spherical harmonics for intrinsic image decomposition," in *Proceedings of ICCV*, 2019, pp. 7820–7829.

[26] Q. Fan, J. Yang, G. Hua, B. Chen, and D. Wipf, "Revisiting deep intrinsic image decompositions," in *Proceedings of CVPR*, 2018, pp. 8944–8952.

[27] Z. Li and N. Snavely, "Cgintrinsics: Better intrinsic image decomposition through physically-based rendering," in *Proceedings of ECCV*, 2018, pp. 371–387.

[28] J. Shi, Y. Dong, H. Su, and S. X. Yu, "Learning non-lambertian object intrinsics across shapenet categories," in *Proceedings of CVPR*, 2017, pp. 5844–5853.

[29] B. Kovacs, S. Bell, N. Snavely, and K. Bala, "Shading annotations in the wild," in *Proceedings of CVPR*, 2017, pp. 6998–7007.

[30] J. Zhu, F. Luan, Y. Huo, Z. Lin, Z. Zhong, D. Xi,

R. Wang, H. Bao, J. Zheng, and R. Tang, "Learning-based inverse rendering of complex indoor scenes with differentiable monte carlo raytracing," in *SIGGRAPH Asia*, 2022, pp. 1–8.

[31] Z. Li, M. Shafiei, R. Ramamoorthi, K. Sunkavalli, and M. Chandraker, "Inverse rendering for complex indoor scenes: Shape, spatially-varying lighting and SVBRDF from a single image," in *Proceedings of CVPR*, 2020, pp. 2475–2484.

[32] R. Zhu, Z. Li, J. Matai, F. Porikli, and M. Chandraker, "Irisformer: Dense vision transformers for single-image inverse rendering in indoor scenes," in *Proceedings of CVPR*, 2022, pp. 2822–2831.

[33] S. Sengupta, J. Gu, K. Kim, G. Liu, D. W. Jacobs, and J. Kautz, "Neural inverse rendering of an indoor scene from a single image," in *Proceedings of ICCV*, 2019, pp. 8598–8607.

[34] Z. Li, J. Shi, S. Bi, R. Zhu, K. Sunkavalli, M. Hašan, Z. Xu, R. Ramamoorthi, and M. Chandraker, "Physically-Based editing of indoor scene lighting from a single image," in *Proceedings of ECCV*, 2022, pp. 555–572.

[35] Z. Wang, J. Philion, S. Fidler, and J. Kautz, "Learning indoor inverse rendering with 3d spatially-varying lighting," in *Proceedings of ICCV*, 2021, pp. 12 538–12 547.

[36] E. Garces, C. Rodriguez-Pardo, D. Casas, and J. Lopez-Moreno, "A survey on intrinsic images: Delving deep into lambert and beyond," *International Journal of Computer Vision*, vol. 130, no. 3, pp. 836–868, 2022.

[37] F.-E. Wang, Y.-H. Yeh, M. Sun, W.-C. Chiu, and Y.-H. Tsai, "BiFuse: Monocular 360 depth estimation via bi-projection fusion," in *Proceedings of CVPR*, June 2020.

[38] H.-L. Shen and Q.-Y. Cai, "Simple and efficient method for specularity removal in an image," *Applied Optics*, vol. 48, no. 14, pp. 2711–2719, 2009.

[39] M.-A. Gardner, Y. Hold-Geoffroy, K. Sunkavalli, C. Gagne, and J.-F. Lalonde, "Deep parametric indoor lighting estimation," in *Proceedings of ICCV*, 2019.

[40] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Proceedings of NeurIPS*, 2021.

[41] G. Fu, Q. Zhang, and C. Xiao, "Towards high-quality intrinsic images in the wild," in *Proceedings of ICME*, 2019, pp. 175–180.

[42] L. Xu, C. Lu, Y. Xu, and J. Jia, "Image smoothing via L0 gradient minimization," *ACM Transactions on Graphics*, pp. 1–12, 2011.

[43] T. Goldstein and S. Osher, "The split bregman method for L1-regularized problems," *SIAM Journal on Imaging Sciences*, vol. 2, no. 2, pp. 323–343, 2009.

[44] J. Zheng, J. Zhang, J. Li, R. Tang, S. Gao, and Z. Zhou, "Structured3d: A large photo-realistic dataset for structured 3d modeling," in *Proceedings of ECCV*, 2020, pp. 519–535.

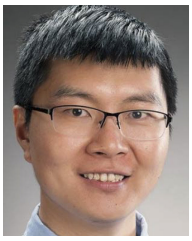
[45] I. Armeni, S. Sax, A. R. Zamir, and S. Savarese, "Joint 2d-3d-semantic data for indoor scene understanding," *arXiv preprint arXiv:1702.01105*, 2017.

[46] H. Weber, D. Prévost, and J.-F. Lalonde, "Learning to estimate indoor lighting from 3d objects," in *Proceedings of 3DV*, 2018, pp. 199–207.

[47] M.-A. Gardner, K. Sunkavalli, E. Yumer, X. Shen, E. Gambaretto, C. Gagné, and J.-F. Lalonde, "Learning to predict indoor illumination from a single image," *ACM Transactions on Graphics.*, vol. 36, no. 6, 2017.

Rong-Kai Xu received the B.S. degree from Beijing Institute of Technology, Beijing, China, in 2022. He is currently a graduate student at the School of Computer Science, Beijing Institute of Technology, Beijing, China. His research interest is computer graphics.

Lei Zhang received the B.S. and Ph.D. degrees from Zhejiang University, Hangzhou, China, in 2004 and 2009, respectively. He is currently a Professor at the School of Computer Science, Beijing Institute of Technology, Beijing, China. His research interests include image and video processing, computer graphics, *etc*. He is a member of IEEE.



Fang-Lue Zhang received the bachelor's degree from Zhejiang University, Hangzhou, China, in 2009, and the doctoral degree from Tsinghua University, Beijing, China, in 2015. He is currently a Senior Lecturer with the Victoria University of Wellington, New Zealand. His research interests include image and video editing, computer vision, and computer graphics. He is a member of IEEE. He received the Victoria Early Career Research Excellence Award in 2019 and Marsden Fast-Start Grant in 2020.