Unsupervised Intrinsic Image Decomposition with LiDAR Intensity

Shogo Sato[1], Yasuhiro Yao[1], Taiga Yoshida[1],
Takuhiro Kaneko[2], Shingo Ando[1], Jun Shimamura[1]
[1]NTT Human Informatics Laboratories, [2]NTT Communication Science Laboratories
{shogo.sato.wv, taiga.yoshida.ry, jun.shimamura.ec}@hco.ntt.co.jp,
yao-yasuhiro@g.ecc.u-tokyo.ac.jp, ando@info.shonan-it.ac.jp

Abstract

Intrinsic image decomposition (IID) is the task that decomposes a natural image into albedo and shade. While IID is typically solved through supervised learning methods, it is not ideal due to the difficulty in observing ground truth albedo and shade in general scenes. Conversely, unsupervised learning methods are currently underperforming supervised learning methods since there are no criteria for solving the ill-posed problems. Recently, light detection and ranging (LiDAR) is widely used due to its ability to make highly precise distance measurements. Thus, we have focused on the utilization of LiDAR, especially LiDAR intensity, to address this issue. In this paper, we propose unsupervised intrinsic image decomposition with LiDAR intensity (IID-LI). Since the conventional unsupervised learning methods consist of image-to-image transformations, simply inputting LiDAR intensity is not an effective approach. Therefore, we design an intensity consistency loss that computes the error between LiDAR intensity and gray-scaled albedo to provide a criterion for the ill-posed problem. In addition, LiDAR intensity is difficult to handle due to its sparsity and occlusion, hence, a LiDAR intensity densification module is proposed. We verified the estimating quality using our own dataset, which include RGB images, LiDAR intensity and human judged annotations. As a result, we achieved an estimation accuracy that outperforms conventional unsupervised learning methods.

1. Introduction

Intrinsic image decomposition (IID) is the task that aims to decompose a natural image into an illumination-invariant component (albedo) and an illumination-variant component (shade), and contributes to high level computer vision tasks such as relighting and scene understanding. Research on decomposing a natural image has a long history, beginning with the proposal of the Retinex theory [19] and IID [2]. Focusing on Lambertian scenes, decomposition of a natural image I is expressed as follows.

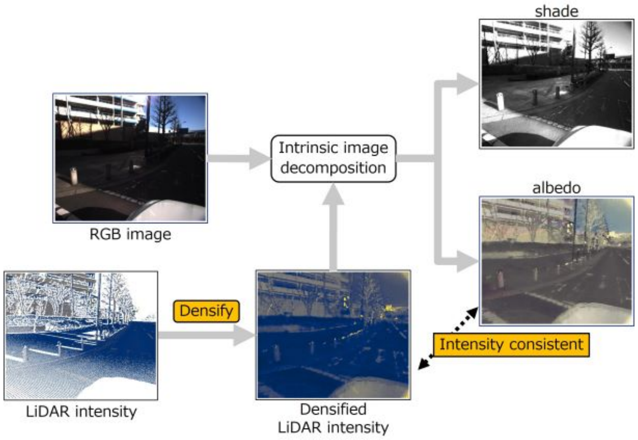


Figure 1. Our proposed approach (IID-LI) is unsupervised intrinsic image decomposition utilizing LiDAR intensity. We densified LiDAR intensity to be robust for LiDAR sparsity or occlusions by a LiDAR intensity densification module. In addition, we designed an intensity consistency loss to provide a criterion for the albedo in IID of ill-posed problems.

$$I = R \cdot S, \tag{1}$$

where, R and S denote albedo and shade, respectively. "$\cdot$" represents a channel-wise multiplication. To solve the ill-posed problem, some researchers assumed that sharp and smooth color variation are caused by albedo and shade change, respectively [12, 19, 41, 44]. As other methods, IID was performed by defining and minimizing energy based on the assumptions such as albedo flatness [3, 4]. Moreover, since shades depend on object geometry, IID methods with a depth map were also proposed [8, 15, 22]. With the development of deep learning, supervised learning methods began to be used for IID [9, 29, 32–34, 45, 46]. Due to the difficulty of observing ground truth albedo and shade in a practical scenario, supervised learning methods are typically either small [12], synthetic [6, 7, 24] or sparsely annotated [3]. Hence, these supervised learning methods are not ideal for IID in observed data. To address this issue, a few semi-supervised [14, 42] and unsupervised [25, 28, 30, 38] learning methods are proposed. However, these methods are currently underperforming supervised learning methods due to the lack of criteria for solving ill-posed problems by only using image and depth.

In recent years, light detection and ranging (LiDAR), which accurately measures the distance to objects, is widely used. LiDAR usually obtains reflectance intensity (LiDAR intensity) as well as object distance. Since albedo is the proportion of the incident light and reflected light, LiDAR intensity utilization as a criterion for albedo helps to solve the ill-posed problem.

In this paper, we propose unsupervised intrinsic image decomposition with LiDAR intensity (IID-LI). The brief flow of IID-LI is depicted in Fig. 1. Since the conventional unsupervised learning methods consist of image-to-image transformations based on variational autoencoder (VAE) [17] , it is not effective to simply input LiDAR intensity. Thus, we design an intensity consistency loss that computes the error between LiDAR intensity and gray-scaled albedo to provide a criterion for the ill-posed problem of decomposing a single image. In addition, LiDAR intensity is difficult to handle due to its sparsity and occlusion, hence, LiDAR intensity densification (LID) module is proposed. The novelty of the LID module lies in the simultaneous convolution of sparse data (LiDAR intensity) and dense data of different modality (RGB image). Then, we verified the estimating quality with our own dataset that combines RGB images and LiDAR intensities in outdoor scenes. In summary, our contributions are as follows.

- We propose LiDAR intensity utilization for intrinsic image decomposition (IID), and an architecture of unsupervised intrinsic image decomposition with LiDAR intensity (IID-LI).
- We design an intensity consistency loss to provide a criterion for the ill-posed problem of decomposing a single image.
- We propose a LiDAR intensity densification (LID) module based on deep image prior (DIP) to be robust for LiDAR sparsity or occlusions.
- We create a publicly available dataset for evaluating IID quality with LiDAR intensity. [1]

The rest of the paper is organized as follows. Sec. 2 and Sec. 3 describe related works and baseline methods, respectively. Sec. 4 explains our proposed method. The details of the experiment and experimental results are described in Sec. 5 and Sec. 6, respectively. Finally, a summary of the research is given in Sec. 7.

2. Related work

In this section, we briefly summarize related works on IID and LiDAR intensity utilization.

2.1. Intrinsic image decomposition (IID)

Optimized based methods. IID represents an ill-posed problem that decomposes a single image into albedo and shade. To address this issue, Land et al. [19] proposed a prior that accounts for sharp and smooth luminance variations induced by albedo and shade changes, respectively. Grosse et al. [12] improved estimation accuracy by considering hue variance as well as luminance. In addition, several priors have been proposed to enhance the estimation accuracy, including the piecewise constant in albedo [1, 26], sparse and discrete values of albedo [35, 37, 39], and similar shade among neighboring pixels [10]. Bell et al. [3] formulated and minimized energy function based on these priors. To achieve edge-preserving smoothing, a combination of global and local smoothness based on superpixels was proposed [4]. Although these handcrafted priors are reasonable in small images, they are insufficient for more complex scenarios, such as outdoor scenes.

Supervised learning methods. With the development of deep learning, supervised learning methods began to be used for IID. Narihira et al. [32] first applied supervised learning methods to IID. In addition, IID was performed by learning the relative reflection intensity of each patch extracted from the image [45]. Nestmeyer et al. [34] directly estimated per-pixel albedo by trained convolutional neural network (CNN). Fan et al. [9] designed a loss function for a universal model, which works on both fully-labeled and weakly labeled datasets. Recently, many researchers have trained supervised learning models based on albedo and shade from synthetic data [29, 46], since observing the ground truth albedo and shade in a practical scenario is difficult. However, the estimation accuracy for observed data may be limited by the gap between the synthetic and observed data.

Unsupervised learning methods. In these days, semi-supervised and unsupervised learning methods have begun to be used in IID. Janner et al. [14] suggested semi-supervised learning methods that utilize a few labeled data for training and transfer to other unlabeled data. Most existing unsupervised learning methods such as Li et al. [25] require a series of images or multi-view images. Liu et al. [28] proposed unsupervised single image intrinsic image decomposition (USI^{3}D), which is an unsupervised learning method from a single image. USI^{3}D outperforms state-of-the-art unsupervised learning methods for IID. However,

[1] Our dataset are publicly available at https://github.com/ntthilab-cv/NTT-intrinsic-dataset

these unsupervised learning methods are currently underperforming supervised learning methods due to the lack of criteria for solving ill-posed problems. In addition, it is difficult for these methods to discriminate between cast shadows and image textures when only using an image and depth. Therefore, we propose the utilization of LiDAR intensity, which is independent of sunlight conditions, cast shadow, and shade.

Datasets for intrinsic image decomposition. MIT Intrinsics, a dataset of image decomposed data on 16 real objects, was published by Grosse et al. [12] as a dataset for IID. Since MIT Intrinsics is small for training deep learning models, synthetic data are widely used. Thus, Butler et al. [6] collected an MPI Sintel dataset, which consisted of synthetic data that included albedo, depth, and optical flow. In addition, synthetic datasets such as the ShapeNet [7] and the CGIntrinsics [24] were also published. The free supervision from video games (FSVG) dataset [18], which was extracted from video games and contains a large number of images in outdoor settings with albedo. On the other hand, Bell et al. [3] published the IIW dataset in real scenes with large number of sparse annotations. Conventionally, datasets with LiDAR intensity and annotations for IID did not exist. To validate the utility of LiDAR intensity for IID, we have created a publicly available dataset that includes RGB images, LiDAR intensity, and IID annotations.

2.2. LiDAR intensity utilization for computer vision

A LiDAR calculates the object distance from the time lags between irradiating the laser and detecting the reflected light. In this process, LiDAR intensity, which is the return strength of a irradiated laser beam, is also obtained. LiDAR intensity is intrinsic to the object surface and is therefore independent of sunlight conditions, cast shadow, and shade. Thus, Guislain et al. [13] performed a shadow detection based on the un-correlation between color and LiDAR intensity. In addition, homogeneous regions are extracted with LiDAR intensity and elevation for cast shadow detection [27]. LiDAR intensity is also utilized for hyper-spectral data correction [5, 36] and object recognition [16, 20, 31]. As mentioned above, LiDAR intensity utilization has the potential to separate scene illumination from albedo, even in unsupervised manner.

3. Baseline method

In this section, we describe the theory of USI^{3}D [28], which is the basis of our proposed method. USI^{3}D was selected as the baseline of IID-LI since it is state-of-the-art unsupervised intrinsic image decomposition method. USI^{3}D [28] is based on VAE [17] with generative adversarial net [11] (VAEGAN) [21]. In USI^{3}D, a sample $X \in \{\text{input } I, \text{ albedo } R, \text{ or shade } S\}$ is decomposed into domain variant (z_X) and in-variant component (c_X). The domain variant component refers to a feature that is unique to each of I, R, and S such as colors, while the domain in-variant component is a common feature of them such as edges. VAE consists of an encoder E^p_X for z_X (prior code encoder), an encoder E^c_X for c_X (content encoder), and a generator G_X from z_X and c_X into a reconstructed image. The image-to-image transformation from I to estimated albedo $R(I)$ and shade $S(I)$ requires the estimation of the domain variant components of albedo and shade $(z_{R(I)}, z_{S(I)})$ corresponding to I. First, the deviation of encoded contents I, $R(I)$ and $S(I)$ is calculated.

$$\mathcal{L}^{\text{cnt}} = |c_{R(I)} - c_I| + |c_{S(I)} - c_I|, \tag{2}$$

where $c_{R(I)}$ and $c_{S(I)}$ are the encoded content of $R(I)$ and $S(I)$, respectively. Second, Kullback-Leibler divergence loss is used to constrain the $z_{R(I)}$ and $z_{S(I)}$ in the albedo prior domain z_R and shade prior domain z_S, respectively.

$$\begin{aligned}\mathcal{L}^{\text{KL}} = \mathbb{E}[\log p(z_{R(I)}) - \log q(z_R)]\\ + \mathbb{E}[\log p(z_{S(I)}) - \log q(z_S)].\end{aligned} \tag{3}$$

On the other hand, VAE computes $\mathcal{L}^{\text{img}}$ and $\mathcal{L}^{\text{pri}}$ to reconstruct the input image and prior code, respectively.

$$\mathcal{L}^{\text{img}} = \sum_{x \in I \text{orRorS}} |G_x(E^c_x(x), E^p_x(x)) - x|, \tag{4}$$

$$\mathcal{L}^{\text{pri}} = \sum_{x \in I \text{orRorS}} |E^p_x(G_x(c_x, z_x)) - z_x|. \tag{5}$$

In addition, adversarial loss $\mathcal{L}^{\text{adv}}$ is computed so that the generated image fits the target domain.

$$\begin{aligned}\mathcal{L}^{\text{adv}}_R = \log(1 - D_R(R(I))) + \log(D_R(R))\\ + \log(1 - D_S(S(I))) + \log(D_S(S)),\end{aligned} \tag{6}$$

where D_R and D_S are discriminators for albedo and shade domain. The IID is based on physical reconstruction and computes the product of albedo and shade to match the input image.

$$\mathcal{L}^{\text{phy}} = |I - R(I) \cdot S(I)|. \tag{7}$$

For a piece-wise constant, the smooth loss is calculated as follows,

$$\mathcal{L}^{\text{smooth}} = \sum_{i=1}^{N} \sum_{j \in N(i)} v_{i,j} |\log x^i_R - \log x^j_R|_1, \tag{8}$$

$$v_{i,j} = \exp\left(-\frac{1}{2}(\vec{f_i} - \vec{f_j})^T \Sigma^{-1} (\vec{f_i} - \vec{f_j})\right), \tag{9}$$

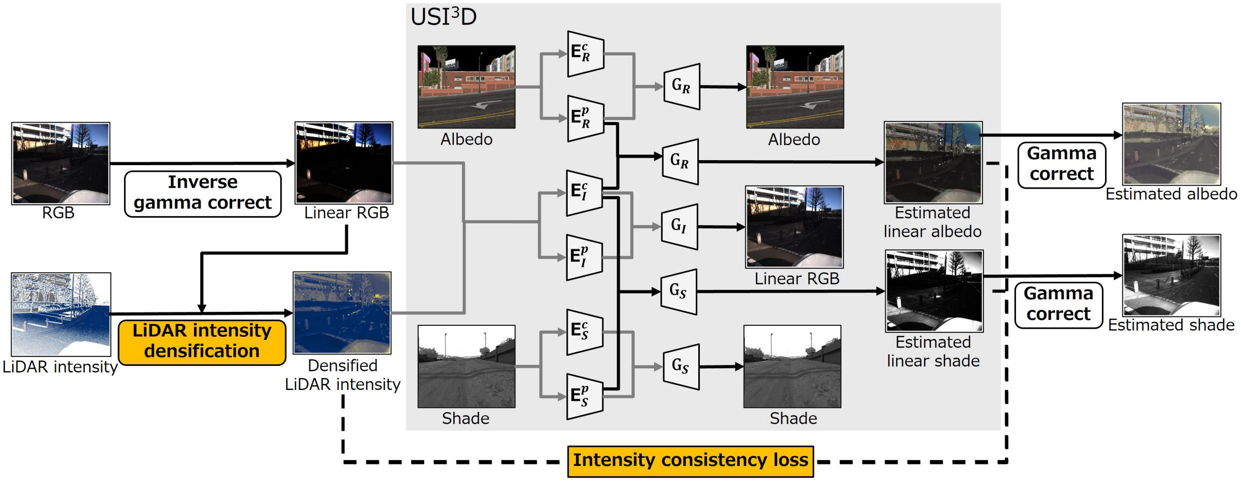


Figure 2. The proposed architecture of IID-LI. Given an RGB image and LiDAR intensity, we convert natural images I into albedo R and shade S domains. The input RGB image is transformed by inverse gamma correction to linearize the image values. Next, a LiDAR intensity densification module is used for robustness against LiDAR sparsity or occlusions. Intensity consistency loss is implemented to provide a criterion for IID. The content encoder, prior encoder, and the decoder are denoted by E_x^c, E_x^p and G_x, respectively.

.

where N and $\vec{f_i}$ are nearest neighbor pixels and i^{th} feature vectors, respectively. The feature vectors are composed of pixel position, image intensity and chromaticity. USI^{3}D optimizes a weighted sum of the above seven types of losses.

$$\min_{E,G,f}\max_{D}(E,G,f,D)=\mathcal{L}^{\text{adv}}+\lambda_1\mathcal{L}^{\text{cnt}}+\lambda_2\mathcal{L}^{\text{KL}}$$
$$+\lambda_3\mathcal{L}^{\text{img}}+\lambda_4\mathcal{L}^{\text{pri}}+\lambda_5\mathcal{L}^{\text{phy}}+\lambda_6\mathcal{L}^{\text{smooth}}. \quad (10)$$

4. Proposed method (IID-LI)

4.1. LiDAR intensity densification module

When utilizing LiDAR intensity for IID, the sparsity or occlusion may cause precision degradation. Thus, we demand to densify the LiDAR intensity before or inside the IID model. In general, dense LiDAR intensity without occlusion corresponding to the supervised data does not exist, and the density of LiDAR intensity is sometimes quite low. Therefore, it is difficult to letting the network manage it implicitly in an end-to-end manner, and we selected a separate model; LID module based on DIP [40]. DIP is one of the fully unsupervised learning method used for denoising and inpainting. In the original DIP, for an image x_0 to be cleaned, a noise map z is input to a deep neural network $f_\theta()$, where θ represents the deep neural network parameters. Then, energy optimization is performed to satisfy Eq. (11).

$$\theta^*=\arg\min_{\theta} E(f_\theta(z);x_0). \quad (11)$$

Based on the difference in learning speed between the noise component and the image component, a clean image is generated by stopping the iteration in the middle of the process. To apply the original DIP directly to LiDAR intensity densification, we substitute LiDAR intensity for x_0 and optimize Eq. (12).

$$\theta^*=\arg\min_{\theta} E(f_\theta(z);m_{\text{L}}x_0), \quad (12)$$

where m_{L} is LiDAR mask which is 1 for pixels with the observed LiDAR intensity and 0 otherwise. However, original DIP tend to blur in sparse regions due to its inconsideration of RGB images. Thus, we input [RGB image, LiDAR intensity]^T and $[m_{\text{R}}, m_{\text{L}}]^T$ for x_0 and m_{L}, respectively. m_{R} has the same shape as the RGB image and all pixel values are 1. By simultaneous convolution of LiDAR intensity and RGB image, LiDAR intensity is densified while considering image edges and brightness. Although DIP is also used in depth completion in multi-view stereo [43], the data, objectives, and loss are all different from our model.

4.2. Intensity consistency loss

The estimation accuracy of IID is highly dependent on the ability to decompose images into domain variant and in-variant components. Utilizing LiDAR intensity alone does not adequately enable the network to learn domain dependence. Thus, we design an intensity consistency loss that computes the error between LiDAR intensity and grayscaled albedo to provide a criterion for letting the network to learn domain dependence efficiently. The first term is the loss function when LiDAR intensity L and the luminance

Method	Train Input	Test Input	Learning	Supervised
Baseline R	single image	single image	No	-
Baseline S	single image	single image	No	-
Retinex [12]	single image	single image	No	-
Color Retinex [12]	single image	single image	No	-
Bell et al. [3]	single image	single image	No	-
Bi et al. [4]	single image	single image	No	-
Revisiting [9]	single image	single image	Yes	Yes
IIDWW [25]	image sequence	single image	Yes	No
UidSequence [23]	a time-varying image pair	single image	Yes	No
USI^{3}D [28]	single image	single image	Yes	No
ours	single image + LiDAR intensity	single image + LiDAR intensity	Yes	No

Table 1. A list of comparison methods, their respective data and training categories.

of the estimated albedo $F(R(I))$ are correlated, where $F()$ is the function to convert from RGB image to gray scale. In the case where LiDAR intensity and albedo are correlated, the image divided by LiDAR intensity is correlated to the shade from Eq. (1). Thus, the second term represents the loss function when $F(I)/L$ and the luminance of the estimated shade $F(S(I))$ are correlated.

$$\begin{aligned}\mathcal{L}^{\text{int}} &= |F(R(I)) - s_1 L - b_1| \cdot m_{\text{L}} \\ &+ |F(S(I)) - s_2(F(I)/L) - b_2| \cdot m_{\text{L}}, \end{aligned} \tag{13}$$

where m_{L} is the mask, which is 1 for pixels with densified LiDAR intensity and 0 otherwise. In addition, s_x and b_x, where $x \in 1, 2$, are trainable parameters for adjusting the scale and bias of LiDAR intensity, respectively. In summary, IID-LI optimizes the loss function in Eq. (14).

$$\begin{aligned}\min_{E,G,f} \max_{D}(E, G, f, D) = \mathcal{L}^{\text{adv}} + \lambda_1 \mathcal{L}^{\text{cnt}} + \lambda_2 \mathcal{L}^{\text{KL}} \\ + \lambda_3 \mathcal{L}^{\text{img}} + \lambda_4 \mathcal{L}^{\text{pri}} + \lambda_5 \mathcal{L}^{\text{phy}} + \lambda_6 \mathcal{L}^{\text{smooth}} + \lambda_7 \mathcal{L}^{\text{int}}\end{aligned} \tag{14}$$

As with the original paper [28], we set $\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5, \lambda_6$ as 10.0, 0.1, 10.0, 0.1, 5.0, and 1.0, respectively. In addition, we set λ_7 as 20.0 in this paper.

4.3. Network architecture

The overview of our proposed method is depicted in Fig. 2. First, LiDAR intensity is densified by LID module since LiDAR points are usually sparse and have occlusion. Then, a RGB image and a densified-LiDAR intensity are given to USI^{3}D-based model with the intensity consistency loss. Since images are generally gamma-corrected for human visibility, the observed brightness is converted nonlinearly. Thus, to compare the images linearly with LiDAR intensity, an inverse gamma correction was performed before input to IID-LI, and gamma correction was performed on the output images.

5. Experimental setups

5.1. Data preparation

In this paper, we attempt to utilize LiDAR intensity for IID. However, no public data with both LiDAR data and annotations for IID exists. Therefore, we employed our own dataset consisting of images, LiDAR data, and annotation. The data was collected using a mobile mapping system (MMS) equipped with RGB camera, LiDAR, global navigation satellite system (GNSS) and an inertial measurement unit (IMU). For LiDAR scanning, the ZF profiler which features 0.0009 degree angular resolution, 1.02 million points per second, 360-degree field of view and a maximum range of 120 meters. The data was recorded at Shinagawa, Yokohama, and Mitaka in Japan.

5.2. Annotation

The annotation in this study follows the conventional method [3]. First, we extracted 100 samples from the obtained dataset for sparse annotation. We then utilized Poisson disk sampling with a minimum radius of 7% of image size to sample image points, and the points with over and under-saturation or those around the edge were removed in the same manner as Bell et al [3]. Finally, Delaunay triangulation was performed to generate edges, resulting in 91 ± 24 pairs per image as sparse sampling. Dense sampling data was generated using Poisson disk sampling with a minimum radius 3% of image size, resulting in 566 ± 143 pairs per image.

In this study, 10 annotators, who understand the concept of albedo, performed the annotation. For each pair, 5 out of 10 annotators answered the following three questions.

- Do the two points have the same albedo intensity?
- If not, does the darker point have a darker surface albedo intensity?

Method	trained dataset	WHDR	precision	recall	F-score
Baseline R	-	0.531	0.393	0.445	0.306
Baseline S	-	0.185	0.431	0.340	0.314
Retinex [12]	-	0.187	0.496	0.455	0.469
Color Retinex [12]	-	0.187	0.496	0.455	0.470
Bell et al. [3]	-	0.213	0.467	0.463	0.457
Bi et al. [4]	-	0.283	0.462	0.522	0.466
Revisiting [9]	IIW	**0.181**	**0.575**	0.485	0.499
IIDWW [25]	BIGTIME	0.375	0.418	0.483	0.397
UidSequence [23]	SUNCG-II	0.372	0.405	0.453	0.395
USI^{3}D [28]	IIW + CGIntrinsics	0.347	0.428	0.497	0.418
USI^{3}D (ours) [28]	ours + FSVG	0.287	0.444	0.504	0.446
Ours (without LID)	ours + FSVG	0.283	0.459	0.530	0.467
Ours (without $\mathcal{L}^{\mathrm{int}}$)	ours + FSVG	0.330	0.426	0.483	0.421
Ours	ours + FSVG	0.227	0.517	**0.591**	**0.521**

Table 2. Numerical comparison with our dataset for **all** (E=9411, D=2554, L=661) annotation points.

- How confident (Definitely, Probably, Guessing) are you in your judgment of the above?

$a_{i,j}$ is the i^{th} annotator judgement for j^{th} question. The $a_{i,1}$ and $a_{i,2}$ value +1 for "yes" and -1 for "no" for the first and second questions, respectively. The $a_{i,3}$ values 1.0, 0.8 and 0.3 for "Definitely", "Probably" and "Guessing", respectively. In this study, the judgement j and the weight w for each pair were calculated as following:

$$(J, w) = \begin{cases} (E, A_1) \text{ if } \mathrm{A_1} > 0 \\ (D, A_2) \text{ if } \mathrm{A_1} \leq 0 \ \mathrm{A_2} > 0 \\ (L, -A_2) \text{ else}, \end{cases} \tag{15}$$

where, $A_1 = \sum_{i=1}^{5} a_{i,1} a_{i,3}$ and $A_2 = \sum_{i=1}^{5} a_{i,2} a_{i,3}$. Note that, annotations are utilized only for quantitative assessment of estimation accuracy.

5.3. Quantitative evaluation metric

For evaluating the estimation accuracy, we defined the threshold differences between the points used in the human judgements.

$$(\hat{J}) = \begin{cases} D \text{ if } \mathrm{R_L}/\mathrm{R_D} > 1 + \delta \\ L \text{ if } \mathrm{R_D}/\mathrm{R_L} > 1 + \delta \\ E \text{ else}, \end{cases} \tag{16}$$

where, R_L and R_D are lighter and darker points of a annotated points pair. Following the conventional studies [3], we set the threshold $\delta = 0.1$. We use four indices: weighted human disagreement rate (WHDR), precision, recall, and F-score. WHDR is calculated as Eq. (17).

$$\mathrm{WHDR}_{\delta}(\mathrm{J}, \mathrm{R}) = \frac{\sum_{\mathrm{k}} \mathrm{w_k} \cdot \mathbf{1}(\mathrm{J_k} \neq \hat{\mathrm{J}}_{\mathrm{k},\delta}(\mathrm{R}))}{\sum_{\mathrm{k}} \mathrm{w_k}}. \tag{17}$$

The human judgement weights are also taken into account when computing precision, recall, and F-score.

5.4. Training details

Our prepared dataset consisted of 10000 samples with RGB images and LiDAR intensities for training. Moreover, we prepared an additional 110 samples for testing, which were combined with a total of 12,626 human judgments. Since IID-LI is based on VAEGAN, both albedo and shade domain data groups are required. Furthermore, to make training fair, the albedo and shade datasets were independent each other. Thus, the albedo dataset was created by extracting 10000 albedo samples from the FSVG dataset [18]. The shade dataset was created with no overlap with the albedo dataset.

6. Evaluation

Initially, we present ten conventional methods for comparison. Next, we provide a quantitative evaluation of these methods based on the metrics of WHDR, precision recall, and F-score.

6.1. Compared methods

In this section, we selected ten compared methods including both optimized-based and learning-based methods listed in Tab. 1. Firstly, "Baseline R" represents an image where all pixel values are 1. Conversely, "Baseline S" decomposes the input image with all the shade pixel values as 1. In addition, Retinex [12], Color Retinex [12], Bell et al. [3], and Bi et al. [4] were selected for optimized-based methods. Finally, supervised learning methods such as Revisiting [9], and unsupervised learning methods such as IIDWW [25], UidSequence [23], and USI^{3}D [28] were

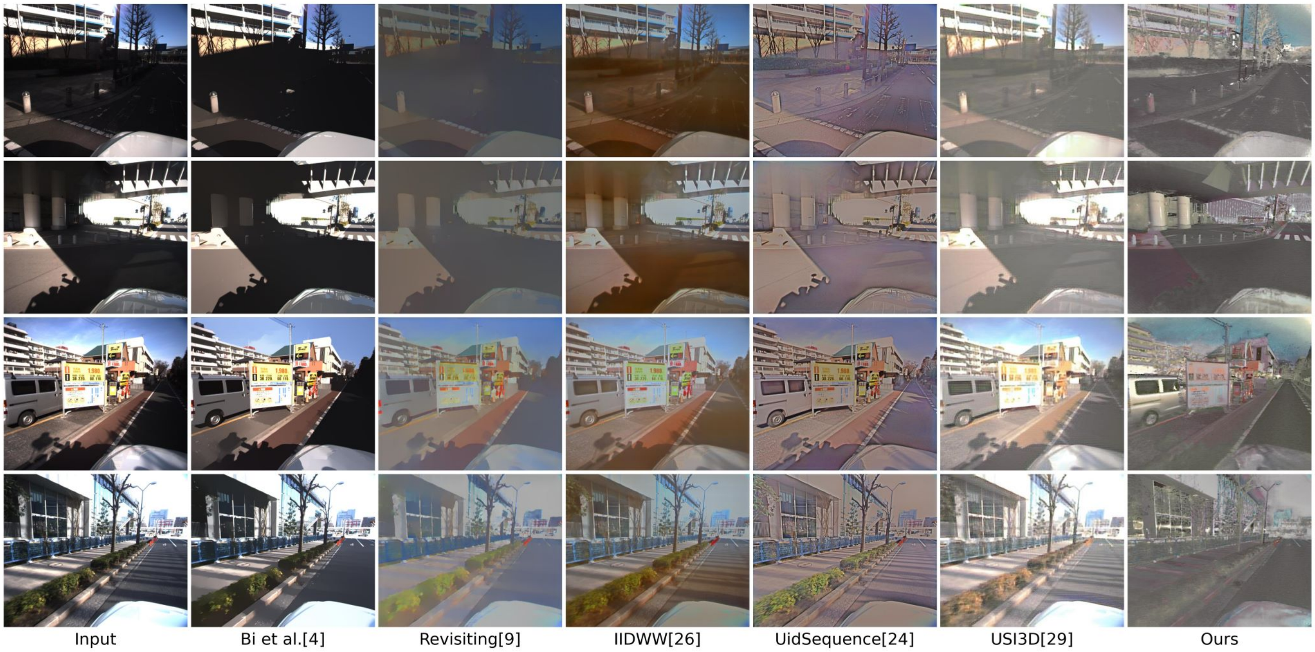


Figure 3. Four examples for IID-LI and compared methods including Bi et al. [4], Revisiting [9], IIDWW [25], UidSequence [23] and USI^{3}D [28]. The conventional methods exhibited noticeable cast shadows, which were reduced with the proposed method.

Method	WHDR	precision	recall	F-score
Baseline R	0.527	0.375	0.440	0.35
Baseline S	0.529	0.361	0.340	0.227
Retinex [12]	0.452	0.523	0.445	0.42
Color Retinex [12]	0.452	0.531	0.445	0.42
Bell et al. [3]	0.446	0.504	0.453	0.414
Bi et al. [4]	0.406	0.561	0.522	0.49
Revisiting [9]	0.428	**0.635**	0.470	0.442
IIDWW [25]	0.464	0.489	0.475	0.417
UidSequence [23]	0.483	0.453	0.450	0.419
USI^{3}D [28]	0.432	0.534	0.500	0.451
USI^{3}D (ours) [28]	0.422	0.539	0.500	0.454
Ours (without LID)	0.410	0.547	0.532	0.534
Ours (without $\mathcal{L}^{\text{int}}$)	0.455	0.513	0.473	0.430
Ours	**0.353**	0.625	**0.596**	**0.602**

Table 3. Numerical comparison with our dataset for **randomly sampled** (E=661, D=661, L=661) annotation points, to eliminate bias in the number of annotations.

employed. For all learning-based methods, publicly available parameters and pre-trained models were used as defaults. Since USI^{3}D is the baseline method, we retrained and validated it with our dataset.

6.2. Qualitative and quantitative evaluation

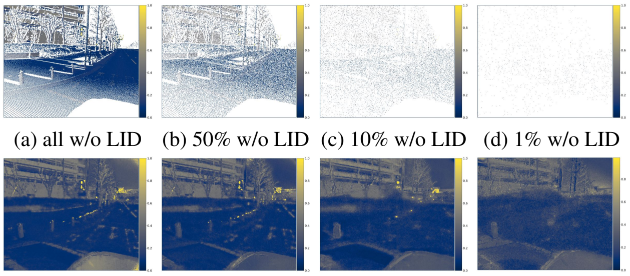
First, we quantitatively evaluated our proposed method and the compared methods. As shown in Tab. 2, our proposed method achieved the state-of-the-art in terms of recall and F-score. However, Revisiting [9] was the best for WHDR and precision. The estimation accuracy of each model may not have been properly evaluated due to the significant bias in the number of "E", "D" and "L" annotations (E=9411, D=2554, L=661). Thus, to eliminate bias in the number of annotations, we randomly sampled the annotations so that each annotation size is the same (E=661, D=661, L=661). Tab. 3 shows the evaluating result for randomly sampled annotations. Our proposed method outperformed other unsupervised learning methods, as shown in Tab. 2 and Tab. 3. Furthermore, it delivered comparable performance to Revisiting [9], which is a supervised learning methods. The visual results are shown in Fig. 3. Cast shadows remained in conventional methods since these methods cannot differentiate between cast shadows and textures. On the other hand, the cast shadows were less noticeable in our proposed method by using LiDAR intensity. However, the images generated by our method are slightly blurred. The image blur is considered to be caused by the calibration errors between the image pixel and LiDAR point, thus addressing this issue is a future work.

Density	WHDR	precision	recall	F-score
all	0.353	0.625	0.596	0.602
50%	0.397	0.574	0.567	0.569
10%	0.437	0.532	0.521	0.524
1%	0.481	0.485	0.480	0.481

Table 4. Ablation study for LiDAR sparsity **with** LID module, when the density of LiDAR intensity was reduced to 50%, 10% and 1% from the original LiDAR intensity.

Density	WHDR	precision	recall	F-score
all	0.410	0.547	0.532	0.534
50%	0.413	0.541	0.524	0.524
10%	0.461	0.493	0.467	0.434
1%	0.480	0.445	0.453	0.418

Table 5. Ablation study for LiDAR sparsity **without** LID module, when the density of LiDAR intensity was reduced to 50%, 10% and 1% from the original LiDAR intensity.


(a) all w/o LID (b) 50% w/o LID (c) 10% w/o LID (d) 1% w/o LID

(e) all w/ LID (f) 50% w/ LID (g) 10% w/ LID (h) 1% w/ LID

Figure 4. An Example of LiDAR intensity with reduced density and the impact of an LID module. (a)-(d): Results of sampling LiDAR points to each density. From left to right, LiDAR points were sampled at 100%, 50%, 10%, and 1%. (e)-(h): Result of densifying LiDAR intensity at each density with LID module.

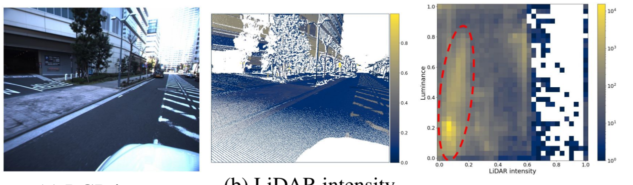
6.3. Ablation study

In our proposed method, we have presented LID module and intensity consistency loss. Thus, the efficacy of these components has been quantified in Tab. 2 and Tab. 3. From the F-score in Tab. 3, the intensity consistency loss contributed the most to estimation accuracy.

This study used relatively high-density LiDAR data, though it can be sparse in certain scenarios. Therefore, we evaluated the estimation accuracy with reduced LiDAR data, and the numerical results for random sampled annotations are listed in Tab. 4. The density of LiDAR intensity was reduced to 50%, 10% and 1% from the original LiDAR intensity. As expected, the higher density of LiDAR intensities indicates higher estimation accuracy, as shown in Tab. 4. Specifically, LiDAR intensity reduced to 1% achieve comparable F-score to the base model, USI^{3}D [28]. As a reference, Tab. 5 shows the estimated results without the LID module, and indicate that the proposed LID module improves the estimation accuracy. Fig. 4 shows an example of LiDAR intensity with reduced LiDAR density, and the effect of LID module. The LiDAR intensity with 1% points were densified successfully in visual, though the estimation accuracy may be limited due to its blurred detail. In a future perspective, the development of a more sparsity-robust model will be interesting.

6.4. Correspondence between albedo and LiDAR intensity

The intensity consistency loss computes the error between LiDAR intensity and gray-scaled albedo. The validity of the intensity consistency loss is discussed in this section. First, an image in which the field of view is almost in shadow, and the corresponding LiDAR intensity is shown in Fig. 5. In addition, each corresponding pixel filled into a 2-dimensional histogram is also shown in Fig. 5 (right). The correlation coefficient between luminance and LiDAR intensity values for this sample was 0.45. Since these maps diverge slightly, we expect that the actual correlations coefficient would be a bit higher. This intensity consistency loss is considered to be effective at least in outdoor scenes with mostly achromatic materials such as concrete and wall surfaces.



(a) RGB image (b) LiDAR intensity (c) Correspondence

Figure 5. Examples of (a) an image in which the field of view is mostly in shadow and (b) the corresponding LiDAR intensities. (c) The correspondence between the luminance and LiDAR intensity are also shown. The red dotted lines in (c) indicate areas of high correlation. The correlation coefficient values 0.45.

7. Conclusion

In this paper, we proposed unsupervised intrinsic image decomposition with LiDAR intensity. We designed an intensity consistency loss and added an LID module for effective LiDAR intensity utilization. As a result, our method outperforms the conventional unsupervised learning methods and is comparable to supervised learning methods with our dataset. In future perspective, we will improve the robustness of LiDAR sparsity, and the registration of image and LiDAR intensity.

References

[1] Jonathan T Barron and Jitendra Malik. Shape, illumination, and reflectance from shading. *IEEE TPAMI*, 37(8):1670–1687, 2014. 2

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

[3] Sean Bell, Kavita Bala, and Noah Snavely. Intrinsic images in the wild. *ACM TOG*, 33(4):1–12, 2014. 1, 2, 3, 5, 6, 7

[4] Sai Bi, Xiaoguang Han, and Yizhou Yu. An l 1 image transform for edge-preserving smoothing and scene-level intrinsic decomposition. *ACM TOG*, 34(4):1–12, 2015. 1, 2, 5, 6, 7

[5] Maximilian Brell, Karl Segl, Luis Guanter, and Bodo Bookhagen. Hyperspectral and lidar intensity data fusion: A framework for the rigorous correction of illumination, anisotropic effects, and cross calibration. *IEEE Transactions on Geoscience and Remote Sensing*, 55(5):2799–2810, 2017. 3

[6] Daniel J Butler, Jonas Wulff, Garrett B Stanley, and Michael J Black. A naturalistic open source movie for optical flow evaluation. In *ECCV*, pages 611–625. Springer, 2012. 1, 3

[7] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 1, 3

[8] Qifeng Chen and Vladlen Koltun. A simple model for intrinsic image decomposition with depth cues. In *ICCV*, pages 241–248, 2013. 1

[9] Qingnan Fan, Jiaolong Yang, Gang Hua, Baoquan Chen, and David Wipf. Revisiting deep intrinsic image decompositions. In *CVPR*, pages 8944–8952, 2018. 1, 2, 5, 6, 7

[10] Elena Garces, Adolfo Munoz, Jorge Lopez-Moreno, and Diego Gutierrez. Intrinsic images by clustering. In *Comput. Graph. Forum*, volume 31, pages 1415–1424. Wiley Online Library, 2012. 2

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *NeurIPS*, volume 27. Curran Associates, Inc., 2014. 3

[12] Roger Grosse, Micah K Johnson, Edward H Adelson, and William T Freeman. Ground truth dataset and baseline evaluations for intrinsic image algorithms. In *ICCV*, pages 2335–2342. IEEE, 2009. 1, 2, 3, 5, 6, 7

[13] M. Guislain, J. Digne, R. Chaine, D. Kudelski, and P. Lefebvre-Albaret. Detecting and Correcting Shadows in Urban Point Clouds and Image Collections. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 537–545, 2016. 3

[14] Michael Janner, Jiajun Wu, Tejas D Kulkarni, Ilker Yildirim, and Josh Tenenbaum. Self-supervised intrinsic image decomposition. *NeurIPS*, 30, 2017. 2

[15] Junho Jeon, Sunghyun Cho, Xin Tong, and Seungyong Lee. Intrinsic image decomposition using structure-texture separation and surface normals. In *ECCV*, pages 218–233. Springer, 2014. 1

[16] Alireza G Kashani and Andrew J Graettinger. Cluster-based roof covering damage detection in ground-based lidar data. *Automation in Construction*, 58:19–27, 2015. 3

[17] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2, 3

[18] Philipp Krähenbühl. Free supervision from video games. In *CVPR*, pages 2955–2964, 2018. 3, 6

[19] Edwin H. Land and John J. McCann. Lightness and retinex theory. *Journal of the Optical Society of America*, 61(1):1–11, Jan 1971. 1, 2

[20] Megan W Lang and Greg W McCarty. Lidar intensity for improved detection of inundation below the forest canopy. *Wetlands*, 29(4):1166–1178, 2009. 3

[21] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1558–1566, New York, New York, USA, 20–22 Jun 2016. PMLR. 3

[22] Kyong Joon Lee, Qi Zhao, Xin Tong, Minmin Gong, Shahram Izadi, Sang Uk Lee, Ping Tan, and Stephen Lin. Estimation of intrinsic image sequences from image+ depth video. In *ECCV*, pages 327–340. Springer, 2012. 1

[23] Louis Lettry, Kenneth Vanhoey, and Luc Van Gool. Unsupervised Deep Single-Image Intrinsic Decomposition using Illumination-Varying Image Sequences. *Comput. Graph. Forum*, 37, 2018. 5, 6, 7

[24] Zhengqi Li and Noah Snavely. Cgintrinsics: Better intrinsic image decomposition through physically-based rendering. In *ECCV*, pages 371–387, 2018. 1, 3

[25] Zhengqi Li and Noah Snavely. Learning intrinsic image decomposition from watching the world. In *CVPR*, pages 9039–9048, 2018. 2, 5, 6, 7

[26] Zicheng Liao, Jason Rock, Yang Wang, and David Forsyth. Non-parametric filtering for geometric detail extraction and material representation. In *CVPR*, pages 963–970, 2013. 2

[27] Xiaoxia Liu, Fengbao Yang, Hong Wei, and Min Gao. Shadow Removal from UAV Images Based on Color and Texture Equalization Compensation of Local Homogeneous Regions. *Remote Sensing*, 14(11):2616, 2022. 3

[28] Yunfei Liu, Yu Li, Shaodi You, and Feng Lu. Unsupervised learning for intrinsic image decomposition from a single image. In *CVPR*, pages 3248–3257, 2020. 2, 3, 5, 6, 7, 8

[29] Jundan Luo, Zhaoyang Huang, Yijin Li, Xiaowei Zhou, Guofeng Zhang, and Hujun Bao. NIID-Net: adapting surface normal knowledge for intrinsic image decomposition in indoor scenes. *IEEE TVCG*, 26(12):3434–3445, 2020. 1, 2

[30] Wei-Chiu Ma, Hang Chu, Bolei Zhou, Raquel Urtasun, and Antonio Torralba. Single image intrinsic decomposition without a single intrinsic image. In *ECCV*, pages 201–217, 2018. 2

[31] Qixia Man, Pinliang Dong, and Huadong Guo. Pixel-and feature-level fusion of hyperspectral and lidar data for urban

land-use classification. *Remote Sensing*, 36(6):1618–1644, 2015. 3

[32] Takuya Narihira, Michael Maire, and Stella X Yu. Direct intrinsics: Learning albedo-shading decomposition by convolutional regression. In *ICCV*, pages 2992–2992, 2015. 1, 2

[33] Takuya Narihira, Michael Maire, and Stella X Yu. Learning lightness from human judgement on relative reflectance. In *CVPR*, pages 2965–2973, 2015. 1

[34] Thomas Nestmeyer and Peter V Gehler. Reflectance adaptive filtering improves intrinsic image estimation. In *CVPR*, pages 6789–6798, 2017. 1, 2

[35] Ido Omer and Michael Werman. Color lines: Image specific color representation. In *CVPR*, volume 2, pages II–II. IEEE, 2004. 2

[36] Frederik Priem and Frank Canters. Synergistic use of LiDAR and APEX hyperspectral data for high-resolution urban land cover mapping. *Remote sensing*, 8(10):787, 2016. 3

[37] Carsten Rother, Martin Kiefel, Lumin Zhang, Bernhard Schölkopf, and Peter Gehler. Recovering intrinsic images with a global sparsity prior on reflectance. *NeurIPS*, 24, 2011. 2

[38] Kouki Seo, Yuma Kinoshita, and Hitoshi Kiya. Deep Retinex Network for Estimating Illumination Colors with Self-Supervised Learning. In *LifeTech*, pages 1–5. IEEE, 2021. 2

[39] Li Shen and Chuohao Yeo. Intrinsic images decomposition using a local and global sparse representation of reflectance. In *CVPR*, pages 697–704. IEEE, 2011. 2

[40] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *CVPR*, pages 9446–9454, 2018. 4

[41] Yair Weiss. Deriving intrinsic images from image sequences. In *ICCV*, volume 2, pages 68–75. IEEE, 2001. 1

[42] Ye Yu and William AP Smith. Inverserendernet: Learning single image inverse rendering. In *CVPR*, pages 3155–3164, 2019. 2

[43] Yinda Zhang and Thomas Funkhouser. Deep depth completion of a single rgb-d image. In *CVPR*, pages 175–185, 2018. 4

[44] Qi Zhao, Ping Tan, Qiang Dai, Li Shen, Enhua Wu, and Stephen Lin. A closed-form solution to retinex with nonlocal texture constraints. *IEEE TPAMI*, 34(7):1437–1444, 2012. 1

[45] Tinghui Zhou, Philipp Krahenbuhl, and Alexei A Efros. Learning data-driven reflectance priors for intrinsic image decomposition. In *ICCV*, pages 3469–3477, 2015. 1, 2

[46] Yongjie Zhu, Jiajun Tang, Si Li, and Boxin Shi. DeRenderNet: Intrinsic Image Decomposition of Urban Scenes with Shape-(In) dependent Shading Rendering. pages 1–11. IEEE, 2021. 1, 2