Unsupervised Single-Image Intrinsic Image Decomposition with LiDAR Intensity Enhanced Training

Shogo Sato, Takuhiro Kaneko, Kazuhiko Murasaki, Taiga Yoshida, Ryuichi Tanida, Akisato Kimura
NTT Corporation
{shg.sato, takuhiro.kaneko, kazuhiko.murasaki, taiga.yoshida, ryuichi.tanida, akisato.kimura}@ntt.com

Abstract

Unsupervised intrinsic image decomposition (IID) is the task of separating a natural image into albedo and shade without ground truth during training. Although a recent model employing light detection and ranging (LiDAR) intensity demonstrated impressive performance, the necessity of LiDAR intensity during inference restricts its practicality. To expand the usage scenario while maintaining the IID quality achieved by using both an image and its corresponding LiDAR intensity, we propose a novel approach that utilizes an image without LiDAR intensity during inference while utilizing both an image and LiDAR intensity during training. Specifically, our proposed model processes an image and LiDAR intensity individually using distinct encoder paths during training, but utilizes only an image-encoder path during inference. Additionally, we introduce an albedo-alignment loss aligning the gray-scale albedo from an image to that from its corresponding LiDAR intensity. LiDAR intensity is not affected by illumination effects including cast shadows, thus albedo-alignment loss transfers the illumination-invariant property of LiDAR intensity to the image-encoder path. Furthermore, we also propose image-LiDAR conversion (ILC) paths that mutually translates the style of an image and LiDAR intensity. IID models translate an image into albedo and shade styles while keeping the image contents, thus it is important to separate the image into contents and style. Trained with pairs of an image and its corresponding LiDAR intensity which share contents but differ in style, the mutual translation in ILC paths improve the accuracy of the separation. Consequently, our model achieves comparable IID quality to the existing model with LiDAR intensity, while utilizing only an image without LiDAR intensity during inference.

1. Introduction

Intrinsic image decomposition (IID) is the process of separating a natural image into an illumination-invariant components (albedo, reflectance) and an illumination-variant components (shade, illumination) within Lambertian scenes. IID provides benefits for various high-level computer vision tasks, such as texture editing [3, 46] and semantic segmentation [2, 60, 61]. The origins of IID can be traced back to early work in computer vision during the 1970s [1, 32], where researchers grappled with the challenge of recovering albedos from shaded images. Since IID is an ill-posed problem that separates a natural image into its albedo and shade, researchers have traditionally addressed this challenge by incorporating a variety of priors, including albedo flatness, shade smoothness [4, 5, 19, 32, 63, 69] and dependence between shade and geometry [11, 27, 35] as energy optimization models. Recently, a notable development has been the emergence of supervised learning models [16, 31, 43, 49, 50, 70, 71, 74],

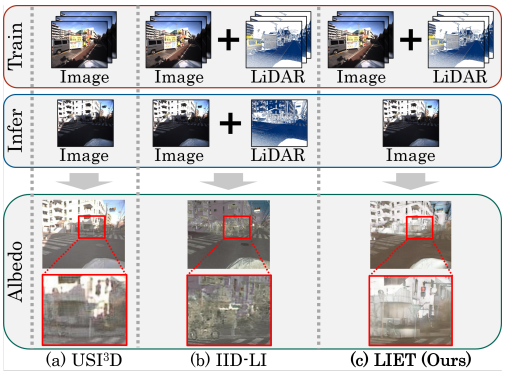


Figure 1. Train/infer schemes and examples of inferred albedos from (a) USI^{3}D [42], (b) IID-LI [54], and (c) LIET (ours). USI^{3}D, utilizing only a single image during training and inference, leaves cast shadows on the inferred albedo. On the other hand, IID-LI utilizes LiDAR intensity during training and inference, making shadows further less noticeable. However, IID-LI has restricted applicability due to the requirement of LiDAR intensity even during inference. LIET utilizes only an image during inference to expand its usage scenarios, and utilizes both an image and LiDAR intensity during training to make shadows less noticeable.

trained on the ground-truth albedo and shade corresponding to an input image with sparsely-annotated datasets [4] or synthetic datasets [9, 10, 39]. However, supervised learning models are not ideal since it is difficult to prepare ground truths by eliminating illumination from images within general scenes. On the other hand, unsupervised learning models [40, 42, 44, 55], that do not utilize ground-truth albedo and shade corresponding to the input image, encountered IID-quality limitations as shown in Fig. 1 (a), particularly in their capacity to reduce cast shadows. More recently, an unsupervised learning model that utilizes light detection and ranging (LiDAR) intensity, called intrinsic image decomposition with LiDAR intensity (IID-LI) [54], has notably enhanced IID quality as shown in Fig. 1 (b). LiDAR intensity refers to the reflected-light strength from object surfaces, and is equivalent to albedo in infrared wavelength. While the LiDAR intensity is effective for IID tasks, the applicability of IID-LI is limited due to its requirement of LiDAR intensity even during inference.

This paper aims to employ only a single image without LiDAR intensity during inference to expand usage scenarios while keeping the high IID quality demonstrated by IID-LI as shown in Fig. 1 (c). To accomplish this objective, we propose *unsupervised single-image intrinsic image decomposition with LiDAR intensity enhanced training (LIET)*. In IID-LI framework, completely-shared model accepts both an image and its corresponding LiDAR intensity as input and processes them simultaneously during both training and inference. On the other hand, LIET is implemented with a *partially-shared model* that processes an image and its corresponding LiDAR intensity individually using an image-encoder path and LiDAR-encoder path, but processes them together during training. The inference from a single image is achieved by utilizing only image-encoder path during inference. Additionally, to enhance the IID quality, we introduce an albedo-alignment (AA) loss aligning the gray-scale albedo from an image to that from its corresponding LiDAR intensity. LiDAR intensity reflects the object-surface properties independent from illumination effects including cast shadows, hence AA loss transfers the illumination-invariant property of LiDAR intensity to the image-encoder path. Due to the lack of hue information in LiDAR intensity, we compare these albedos in gray scale. Furthermore, we also propose image-LiDAR conversion (ILC) paths that mutually translates the style of an image and LiDAR intensity. IID models translate an image into albedo and shade styles while keeping the image contents, thus it is important to separate the image into content and style codes[1]. Due to the shared content but differing styles between an image and its corresponding LiDAR intensity, the ILC paths that mutually translate them facilitate separating the image into content and style codes, enhancing the IID quality.

The performance of LIET is investigated by comparing it with existing IID models including energy optimization models [4, 5, 19], weakly supervised model [16], and unsupervised models [36, 40, 42, 54] in IID quality metrics and image quality assessment (IQA) [17, 30, 58, 64, 67] on inferred albedos. The main contributions of this study are summarized as follows.

- We propose *unsupervised single-image intrinsic image decomposition with LiDAR intensity enhanced training (LIET)* with a *partially-shared model* that processes an image and its corresponding LiDAR intensity individually using an image-encoder path and LiDAR-encoder path, but processes them together during training. The inference from a single image without LiDAR intensity is achieved by utilizing only image-encoder path during inference.
- To enhance the effective utilization of LiDAR intensity, we introduce *albedo-alignment loss* to align the albedo inferred from an image to that from its corresponding LiDAR intensity, and *image-LiDAR conversion (ILC) paths* to translate an image into albedo and shade style while keeping the image contents.
- In terms of IID quality, LIET, which employs only an image during inference demonstrates comparable performance to the existing model, which employs both an image and LiDAR intensity during inference. Additionally, the ablation study demonstrates the effectiveness of each proposed architecture and loss.

2. Related work

This section initially introduces general image-to-image translation (I2I) models that translate an input image from their source domain to the target domain, since IID represents a specific form of I2I that translates an input image from the image domain into albedo and shade domains. Furthermore, we describe the existing unsupervised IID models and examples of LiDAR intensity utilization in this section.

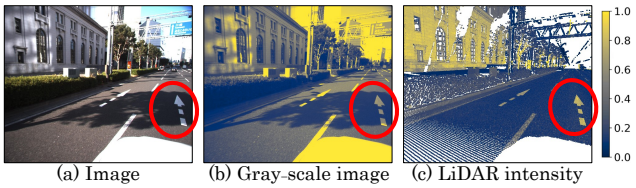
Image-to-image translation. I2I models are designed to translate an input image from their source domain to the target domain. Most of the I2I models rely on generative models include generative adversarial networks (GAN) [18] such as pix2pix [26]. Due to the challenge of acquiring paired images for each domain, CycleGAN [73] was proposed as an I2I model that does not require paired images. Additionally, unsupervised image-to-image translation networks (UNIT) [41] achieved unsupervised I2I by implementing weight sharing within the latent space. Conversely, UNIT and CycleGAN require training as many models as the number of domains to be translated, leading to high computational costs. Thus, StarGAN [14]

[1]Style and contents denote domain-variant component like illuminations, and domain-invariant components like object edges, respectively.

and multimodal unsupervised image-to-image translation (MUNIT) [25] were introduced to translate images into multiple domains using a single model. Additionally, diverse image-to-image translation via disentangled representations (DRIT) [34], which amalgamates the advantages of UNIT and MUNIT, was presented. More recently, diffusion model [23, 52] began to be applied for I2I [13, 37].

Unsupervised IID models. Acquiring ground-truth albedo and shade corresponding to an input image in general scenes presents a considerable challenge, thus necessitating the use of unsupervised learning models that do not depend on such ground truths. To facilitate IID in the absence of ground truths, two primary strategies are employed, one using multiple images captured under different conditions and the other using a single image and synthetic data for albedo and shade domains that do not correspond to the image. As the first strategy, models have been trained using pairs of images under varying illumination conditions [44], as well as sequences of related images [65]. More recent models [6, 7, 21, 48, 57, 66, 68, 72] leveraging the neural radiance field (NeRF) [47] framework from multi-view images have been proposed. Conversely, as the second strategy, USI^{3}D [42] employs an image for IID and the synthetic albedo and shade domain data for ensuring albedo or shade domain likelihood, resulting in the enhancement of the decomposition quality without direct ground truth. In addition, IID-LI [54] has incorporated LiDAR intensity based on USI^{3}D to reduce cast shadows and demonstrated impressive IID performance. However, IID-LI has restricted applicability due to the requirement for LiDAR intensity even during inference. Thus, IID models employing only a single image during inference while keeping as high IID quality as IID-LI are highly desired.

LiDAR intensity utilization. LiDAR is a device for measuring the distance to the object surfaces based on the time of flight from infrared laser irradiation to the reception of reflected light. In addition, LiDAR also captures the intensity of the reflected light from the object surfaces, commonly referred to LiDAR intensity. This LiDAR intensity is unaffected by variations in sunlight conditions or shading while preserving the texture of object surfaces as illustrated in Fig. 2. Thus, LiDAR intensity has the potential for effective utilization in the context of IID tasks. Also, LiDAR intensity is widely applied due to its ability to depict surface properties, for instance, shadow detection [20, 53], hyper-spectral data correction [8, 51], and object recognition [29,33,38,45]. Note that LiDAR and a camera typically operate in distinct wavelength bands; the near-infrared band and the visible light band, respectively. Hence, LiDAR intensity is not to be used directly as for ground-truth albedo. LIET approaches the difference in wavelength by calculating loss with albedo inferred from LiDAR intensity, rather than direct utilization.



(a) Image (b) Gray-scale image (c) LiDAR intensity

Figure 2. Examples of (a) input image, (b) gray-scale image, and (c) its corresponding LiDAR intensity. The red circle indicates the regions with cast shadows and white arrows. The shadow and the white arrow are visible in gray-scale image. LiDAR intensity has no cast shadows while maintaining white arrows, since LiDAR intensity is calculated from the intensity ratio of irradiated and reflected lights, equivalent to an albedo at infrared wavelength.

3. Proposed model (LIET)

3.1. Problem formulation

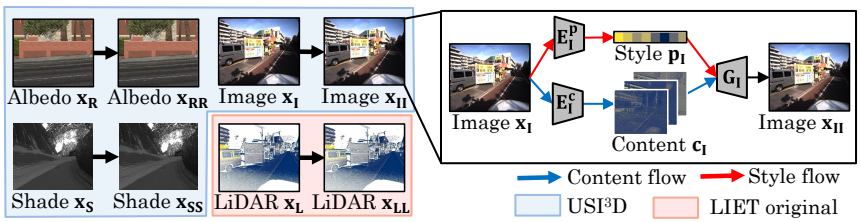
This section describes the problem formulation addressed by LIET framework. Firstly, the inference process employs only a real-world image x_{I} to infer albedo x_{RI} and shade x_{SI}. Meanwhile the training process employs a set of a real-world image x_{I} and its corresponding LiDAR intensity x_{L} to infer albedo x_{RI} and shade x_{SI}. The ground-truth albedo and shade corresponding to the image are not utilized due the difficulty of obtaining ground truths in the real world. Instead of these ground truths, albedo x_{R} and shade x_{S} derived from synthetic dataset that do not corresponding to the image x_{I} are utilized, thereby enabling calculation of the distributions for albedo and shade.

3.2. USI^{3}D architecture

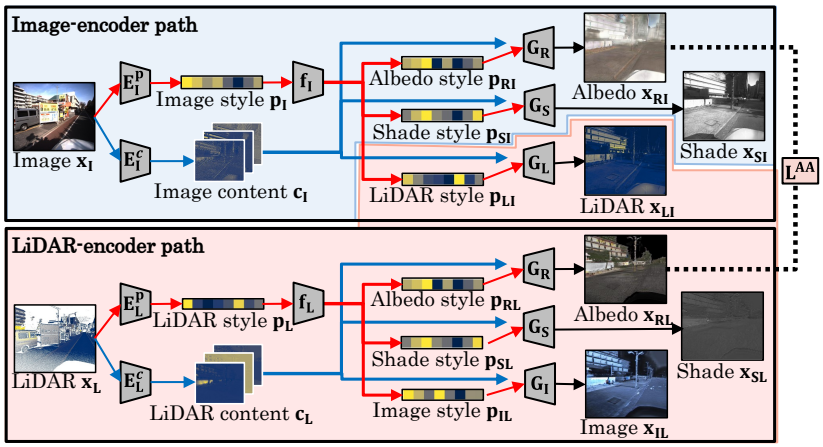
Before introducing the details of LIET, we describe USI^{3}D, which is the baseline of LIET.

Overview. USI^{3}D [42] consists of within-domain reconstruction and cross-domain translation as illustrated in light-blue regions of Fig. 3. The within-domain reconstruction aims to extract features for each domain of image x_{I}, albedo x_{R}, and shade x_{S} by encoders and decoders. The cross-domain translation infers albedo x_{RI} and shade x_{SI} from an input image x_{I}. Simultaneously inferring multiple domains helps extract contents common across domains, hence inferring both albedo and shade leads to enhance IID quality compared to inferring only albedo.

Within-domain reconstruction. As illustrated in Fig. 3 (a), for each domain including image I, albedo R, and shade S, an input x_{X} is fed into style encoder E^p_{X} and content encoder E^c_{X} to derive style code p_{X} and content code c_{X}, respectively, for $X \in \{\mathrm{I}, \mathrm{R}, \mathrm{S}\}$. These codes are input to domain-specific generators G_{X}, reconstructing the inputs within their respective domains x_{XX}. For example, for an image x_{I}, the image-style encoder E^p_{I} and image-content encoder E^c_{I} are utilized to extract image-style code p_{I} and



(a) Within-domain reconstruction



(b) Cross-domain translation

Figure 3. LIET architecture including (a) within-domain reconstruction and (b) cross-domain translation. (a) For each domain (image I, LiDAR intensity L, albedo R, shade S), an input x_{X} is fed into style E^p_{X} and content E^c_{X} encoders to calculate style p_{X} and content c_{X} codes for domain $X \in \{\mathrm{I}, \mathrm{L}, \mathrm{R}, \mathrm{S}\}$. These codes are used at generators G_{X} to reconstruct the inputs within their domains. (b) The image-encoder path accepts an image x_{I} and infers albedo x_{RI}, shade x_{SI}, and LiDAR x_{LI} through image encoders $(E^p_{\mathrm{I}}, E^c_{\mathrm{I}})$, a style mapping function f_{I}, and generators $(G_{\mathrm{R}}, G_{\mathrm{S}}, G_{\mathrm{L}})$. Similarly, the LiDAR-encoder path uses x_{L} to infer albedo x_{RL}, shade x_{SL}, and image x_{IL}. The AA loss $\mathcal{L}^{\mathrm{AA}}$ aligns the gray-scaled albedo from an image to that inferred from the LiDAR intensity to reduce cast shadows.

image-content code c_{I}, respectively. The reconstructed image x_{II} is inferred by image generator $G_{\mathrm{I}}(c_{\mathrm{I}}, p_{\mathrm{I}})$. Analogous processes apply for albedo and shade reconstructions.

Cross-domain translation. For cross-domain translation, an input image x_{I} is processed to infer albedo x_{RI} and shade x_{SI}. First, an input x_{I} is fed into image-style encoder E^p_{I} and image-content encoder E^c_{I} to derive image-style code p_{I} and image-content code c_{I}, respectively. A style mapping function f_{I} then adjusts p_{I} to generate domain-specific style codes for albedo p_{RI} and shade p_{SI}. Subsequently, these style codes, alongside c_{I}, are input into their respective generators $(G_{\mathrm{R}}(c_{\mathrm{I}}, p_{\mathrm{RI}}), G_{\mathrm{S}}(c_{\mathrm{I}}, p_{\mathrm{SI}}))$ to infer the albedo x_{RI} and shade x_{SI}, respectively. Note that all encoders and decoders are shared between within-domain reconstruction and cross-domain translation.

Adversarial training. To improve the translation quality from the source domain into the target domain, the translated images are input into discriminators. Discriminators for albedo D_{R} and shade D_{S} are implemented for adversarial training. For instance, the albedo inferred from image x_{RI} and the albedo domain data x_{R} are provided into the albedo discriminator D_{R} to evaluate the domain likelihood.

3.3. LIET architectures

Overview. LIET is based on USI^{3}D and consists of within-domain reconstruction and cross-domain translation as illustrated in Fig. 3. In the same manner as USI^{3}D, LIET infers albedo x_{RI} and shade x_{SI} from an image x_{I} in image-encoder path. To support the image-encoder path, we introduce LiDAR-encoder path inferring albedo x_{RL} and shade x_{SL} from LiDAR intensity x_{L}. In addition, to transfer the illumination-invariant property of LiDAR intensity to the image-encoder path, we introduce an AA loss $\mathcal{L}^{\mathrm{AA}}$ aligning the albedo inferred from an image x_{RI} to that from its corresponding LiDAR intensity x_{RL}. Since an image and its corresponding LiDAR intensity share the contents but differ in style, we also introduce ILC paths that mutually translate the style of image and LiDAR intensity to support the separation of an image into contents and style. The inference from an image is achieved by utilizing only the image-encoder path during inference, while both paths are utilized during training. Note that all encoders and decoders are shared in within-domain reconstruction, the image-encoder path, and the LiDAR-encoder path. For comparing an image linearly with LiDAR intensity, we convert a sRGB im-

age to linear RGB before inputting it into LIET, and revert the inferred albedo and shade back to sRGB for better human visibility.

Within-domain reconstruction. In the same manner as USI^{3}D, within-domain reconstructions for image, albedo, and shade are implemented. In addition, reconstruction for LiDAR intensity x_{L} is also implemented in LIET as shown in Fig. 3 (a), to extract features of LiDAR intensity. The LiDAR intensity x_{L} is input into LiDAR-style encoder E^p_{L} and LiDAR-content encoder E^c_{L} to calculate the LiDAR-style code p_{L} and LiDAR-content code c_{L}, respectively. Subsequently, LiDAR generator $G_{\mathrm{L}}(c_{\mathrm{L}}, p_{\mathrm{L}})$ is utilized to reconstruct the LiDAR intensity x_{LL}.

Cross-domain translation. As illustrated in the light-blue region of Fig. 3 (b), USI^{3}D infers albedo x_{RI} and shade x_{SI} from an image x_{I}. Within the LIET framework, the objective is to maintain the use of solely an image x_{I} as inputs during inference, while simultaneously leveraging both the image x_{I} and LiDAR intensity x_{L} during training. To this end, alongside the conventional image-encoder path, a LiDAR-encoder path employs LiDAR intensity x_{L} to infer albedo x_{RL} and shade x_{SL}. The LiDAR intensity x_{L} is fed into LiDAR-style encoder E^p_{L} and LiDAR-content encoder E^c_{L} to calculate the LiDAR-style p_{L} and LiDAR-content c_{L} codes, respectively. The LiDAR-style p_{L} is input into the style mapping function f_{L} to infer style codes for albedo p_{RL} and shade p_{SL}. Finally, these style and content codes are fed into generators $(G_{\mathrm{R}}(c_{\mathrm{L}}, p_{\mathrm{RL}}), G_{\mathrm{S}}(c_{\mathrm{L}}, p_{\mathrm{SL}}))$ to infer albedo x_{RL} and shade x_{SL}. This partially-shared model supports the concurrent objectives of effectively leveraging LiDAR intensity during training and maintaining exclusive reliance on an image input during inference.

Image-LiDAR conversion (ILC) paths. As illustrated in the light-blue region of Fig. 3 (b), USI^{3}D separates an image x_{I} into content code c_{I} and style code p_{I} to infer albedo x_{RI} and shade x_{SI}, and this separation process is critical as it directly impacts the IID quality. Since USI^{3}D utilizes the independence datasets for image x_{I}, albedo x_{R} and shade x_{S} without shared content, this separation relies solely on the style information unique to each domain. On the other hand, in this problem formulation, LiDAR intensity x_{L} corresponding to the image x_{I} is also available. This helps us separate an image into content and style codes $(c_{\mathrm{I}}, p_{\mathrm{I}})$ accurately since the image and LiDAR intensity should share the content. LIET incorporates the decoder $G_{\mathrm{L}}(c_{\mathrm{I}}, p_{\mathrm{LI}})$ for inferring the LiDAR intensity in the image-encoder path, thereby, the content code c_{I} and style code p_{I} of the input image are effortlessly separated. Furthermore, to enhance the inference quality of albedo x_{RL} and shade x_{SL} from the LiDAR intensity x_{L}, the image inference path is also added to the LiDAR-encoder path in the same manner as the image to LiDAR intensity path.

Adversarial training. Similar to USI^{3}D, inferred albedo x_{RI} and shade x_{SI} from an image x_{I} is input into albedo discriminator D_{R} and shade discriminator D_{S}, respectively, to improve the translation quality from the source domain into the target domain. In LIET, due to the incorporation of the LiDAR-encoder path, the albedo x_{RL} and shade x_{SL} inferred from LiDAR intensity x_{L} are also input into their respective discriminators $(D_{\mathrm{R}}, D_{\mathrm{S}})$. Furthermore, along with the ILC paths, inferred LiDAR intensity x_{LI} and image x_{IL} are also evaluated by LiDAR-intensity discriminator D_{L} and image discriminator D_{I}, respectively.

3.4. Losses

In this section, we describe the loss functions computed during within-domain reconstruction and cross-domain translation.

Image reconstruction loss $\mathcal{L}^{\mathrm{img}}$. Initially, the input images should be reconstructed after passing through the within-domain reconstruction process; hence image reconstruction loss $\mathcal{L}^{\mathrm{img}}$ is defined in Eq. (1).

$$\mathcal{L}^{\mathrm{img}} = \sum_{X \in \{I,L,R,S\}} |x_{\mathrm{XX}} - x_{\mathrm{X}}|, \tag{1}$$

where x_{II}, x_{LL}, x_{RR}, and x_{SS} are reconstructed images by within-domain reconstruction for image, LiDAR intensity, albedo, and shade, respectively.

Style reconstruction loss $\mathcal{L}^{\mathrm{sty}}$ and content code reconstruction loss $\mathcal{L}^{\mathrm{cnt}}$. Since the reconstructed images should maintain their styles and contents, the style reconstruction loss $\mathcal{L}^{\mathrm{sty}}$ and content code reconstruction loss $\mathcal{L}^{\mathrm{cnt}}$ are defined in Eq. (2) and Eq. (3), respectively.

$$\mathcal{L}^{\mathrm{sty}} = \sum_{X \in \{L,R,S\}} |E^p_{\mathrm{X}}(x_{\mathrm{XI}}) - p_{\mathrm{XI}}| + \sum_{X \in \{I,R,S\}} |E^p_{\mathrm{X}}(x_{\mathrm{XL}}) - p_{\mathrm{XL}}|, \tag{2}$$

$$\mathcal{L}^{\mathrm{cnt}} = \sum_{X \in \{L,R,S\}} |E^c_{\mathrm{X}}(x_{\mathrm{XI}}) - c_{\mathrm{XI}}| + \sum_{X \in \{I,R,S\}} |E^c_{\mathrm{X}}(x_{\mathrm{XL}}) - c_{\mathrm{XL}}|. \tag{3}$$

Adversarial loss $\mathcal{L}^{\mathrm{adv}}$. Moreover, the adversarial loss $\mathcal{L}^{\mathrm{adv}}$ [18] is defined as Eq. (4) to ensure that the image inferred through cross-domain translation aligns with the distribution of the target domain.

$$\begin{aligned}\mathcal{L}^{\mathrm{adv}} = &\sum_{X \in \{L,R,S\}} \log(1 - D_{\mathrm{X}}(x_{\mathrm{XI}})) \\ &+ \sum_{X \in \{I,R,S\}} \log(1 - D_{\mathrm{X}}(x_{\mathrm{XL}})) + \sum_{X \in \{I,L,R,S\}} \log(D_{\mathrm{X}}(x_{\mathrm{X}}))\end{aligned} \tag{4}$$

VGG loss $\mathcal{L}^{\mathrm{VGG}}$. To preserve the object edges and colors of the input image, the distance between the input image and

the inferred albedo within the VGG feature space is computed [12, 56, 62] for the VGG loss $\mathcal{L}^{\mathrm{VGG}}$ [28] in Eq. (5).

$$\mathcal{L}^{\mathrm{VGG}} = |V(x_{\mathrm{I}}) - V(x_{\mathrm{RI}})|, \tag{5}$$

where V is pre-trained visual-perception network such as VGG-19 [56].

KLD loss $\mathcal{L}^{\mathrm{KLD}}$. Additionally, the Kullback-Leibler divergence (KLD) loss $\mathcal{L}^{\mathrm{KLD}}$ is formulated as Eq. (6) to align the distributions of inferred albedo style p_{RI} and shade style p_{SI} from an image with those calculated from synthetic data $(p_{\mathrm{R}}, p_{\mathrm{S}})$ facilitated by a style mapping function.

$$\mathcal{L}^{\mathrm{KLD}} = p_{\mathrm{R}} \cdot \log \frac{p_{\mathrm{R}}}{p_{\mathrm{RI}}} + p_{\mathrm{S}} \cdot \log \frac{p_{\mathrm{S}}}{p_{\mathrm{SI}}}. \tag{6}$$

Physical loss $\mathcal{L}^{\mathrm{phy}}$. Given the assumption of a Lambertian surface in the IID task, the product of albedo and shade is expected to match the input image. Thus physical loss $\mathcal{L}^{\mathrm{phy}}$ is defined in Eq. (7).

$$\mathcal{L}^{\mathrm{phy}} = |x_{\mathrm{I}} - x_{\mathrm{RI}} \cdot x_{\mathrm{SI}}| + |x_{\mathrm{L}} - x_{\mathrm{RL}} \cdot x_{\mathrm{SL}}|. \tag{7}$$

Albedo-alignment loss $\mathcal{L}^{\mathrm{AA}}$. To improve the IID quality, we propose AA loss $\mathcal{L}^{\mathrm{AA}}$ as depicted in Fig. 4, aligning the gray-scale albedo from an image to that inferred from its corresponding LiDAR intensity. Since the albedo inferred from LiDAR intensity x_{RL} is independent of daylight conditions and cast shadows, the IID quality is expected to improve by aligning the albedo inferred from an image x_{RI} to that inferred from LiDAR intensity x_{RL}. Additionally, these albedos are required to compare in gray scale due to the lack of hue in LiDAR intensity. Thus, AA loss $\mathcal{L}^{\mathrm{AA}}$ is defined in Eq. (8) to compute the distance between x_{RI} and x_{RL} in gray scale.

$$\mathcal{L}^{\mathrm{AA}} = |\mathrm{In}(\mathrm{x}_{\mathrm{RI}} \cdot \mathrm{m}_{\mathrm{L}}) - \mathrm{In}(\mathrm{x}_{\mathrm{RL}} \cdot \mathrm{m}_{\mathrm{L}})|, \tag{8}$$

where m_{L} represents the mask denoting the presence of LiDAR intensity values. $\mathrm{In}(\cdot)$ is an instance normalization [59] and gray-scale function. The instance normalization is used to align the scales of x_{RI} and x_{RL}. In addition, a stop gradient is performed on the LiDAR-encoder path side to align x_{RI} to x_{RL}.

In summary, LIET optimizes the loss function in Eq. (9).

$$\begin{aligned}\mathcal{L}^{\mathrm{LIET}} &= \mathcal{L}^{\mathrm{adv}} + \lambda_{\mathrm{img}}\mathcal{L}^{\mathrm{img}} + \lambda_{\mathrm{sty}}\mathcal{L}^{\mathrm{sty}} + \lambda_{\mathrm{cnt}}\mathcal{L}^{\mathrm{cnt}} \\ &+ \lambda_{\mathrm{KLD}}\mathcal{L}^{\mathrm{KLD}} + \lambda_{\mathrm{VGG}}\mathcal{L}^{\mathrm{VGG}} + \lambda_{\mathrm{phy}}\mathcal{L}^{\mathrm{phy}} + \lambda_{\mathrm{AA}}\mathcal{L}^{\mathrm{AA}}.\end{aligned} \tag{9}$$

$\lambda_{\mathrm{img}}, \lambda_{\mathrm{sty}}, \lambda_{\mathrm{cnt}}, \lambda_{\mathrm{KLD}}, \lambda_{\mathrm{VGG}}, \lambda_{\mathrm{phy}}$, and λ_{AA} are hyper parameters for balancing the losses. The effects of each hyper parameter are detailed in the supplementary materials.

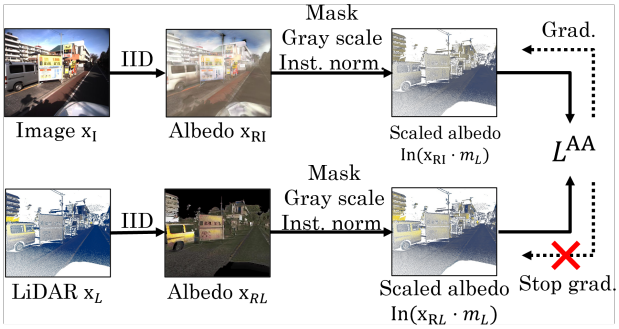


Figure 4. Calculation process of AA loss $\mathcal{L}^{\mathrm{AA}}$. First, albedo from an image x_{RI} and that from LiDAR intensity x_{RL} are computed. Subsequently, these albedos are masked to the points with LiDAR values and then gray scaled. Next, instance normalization (inst. norm.) is performed to align the scales of these albedos, and the distance between these scaled albedos is calculated. A stop gradient is performed on the LiDAR-encoder path side to align x_{RI} to x_{RL} since the LiDAR intensity is independent of sunlight conditions. LiDAR intensity, scaled albedo from the image, and that from LiDAR intensity are represented in a cividis color map due to their gray scale.

4. Experiments

4.1. Experimental setting

Dataset. To facilitate IID using LiDAR intensity during training, we employ the NTT-IID dataset [54], which consists of images, LiDAR intensities, and annotations designed for evaluating IID quality. The NTT-IID dataset prepares 10,000 pairs of images measuring outdoor scenes and LiDAR intensity mapped to these images. Among these pairs, 110 samples have been annotated, yielding a total of 12,626 human judgments. Additionally, we utilized the FSVG dataset [31] as the target domain for albedo and shade. In this paper, we employ the same albedo and shade samples as those used in IID-LI [54]. Note that, the shade samples in FSVG dataset represent in grayscale, shades in this paper do not consider illumination color as following to previous papers [42, 54].

Evaluation metrics. For quantitative evaluation, we employ the following metrics: the weighted human disagreement rate (WHDR), precision, recall, and F-score for all and random sampled annotation[2], following the same methodology as IID-LI [54]. In addition, we evaluate the image quality using five IQA models: MANIQA [64], TReS [17], MUSIQ [30], HyperIQA [58], and DBCNN [67].

Implementation details. In LIET, the style code encoder E^{p}_{X}, content code encoder E^{c}_{X}, generator G_{X}, and discriminator D_{X}, $(X \in \{I, L, R, S\})$, are implemented with the

[2]NTT-IID dataset [54] also provides random sampled annotations to eliminate bias in the number of annotations.

Model	Learning	Random sampled annotation				All annotation			
		F-score(↑)	WHDR(↓)	Precision(↑)	Recall(↑)	F-score(↑)	WHDR(↓)	Precision(↑)	Recall(↑)
Baseline-R [4]	No	0.350	0.527	0.375	0.440	0.306	0.531	0.393	0.445
Baseline-S [4]	No	0.227	0.529	0.361	0.340	0.314	0.185	0.431	0.340
Retinex [19]	No	0.420	0.452	0.523	0.445	0.469	0.187	0.496	0.455
Color Retinex [19]	No	0.420	0.452	0.531	0.445	0.470	0.187	0.496	0.455
Bell et al. [4]	No	0.414	0.446	0.504	0.453	0.457	0.213	0.467	0.463
Bi et al. [5]	No	0.490	0.406	0.561	0.522	0.466	0.283	0.462	0.522
IIDWW [40]	Unsup.	0.417	0.464	0.489	0.475	0.397	0.375	0.418	0.483
UidSequence [36]	Unsup.	0.419	0.483	0.453	0.450	0.395	0.372	0.405	0.453
USI^{3}D [42]	Unsup.	0.454	0.422	0.539	0.500	0.446	0.287	0.444	0.504
IID-LI [54]	Unsup.	0.602	0.353	0.625	0.596	0.521	0.227	0.517	0.591
LIET (ours)	Unsup.	0.607	0.340	0.649	0.601	0.525	0.245	0.500	0.598
Revisiting* [16]	Sup.	0.442	0.428	0.635	0.470	0.499	0.181	0.575	0.485

Table 1. Numerical comparison in IID quality with NTT-IID dataset [54]. Along with the existing paper [54], we evaluated (i) randomly sampled annotation and (ii) all annotations. Red and blue fonts indicate the best and second-best results, respectively. For both randomly sampled annotation and all annotations, LIET (ours) achieves comparable IID quality to IID-LI. In addition, Revisiting* [16] demonstrates the better IID quality in two indices of all annotations. Due to biases in the distribution of all annotations, models that infers flatter albedos are at an advantage in these metrics. Revisiting* [16], assuming local flatness of albedo in its training, demonstrates superior results by aligning with this bias.

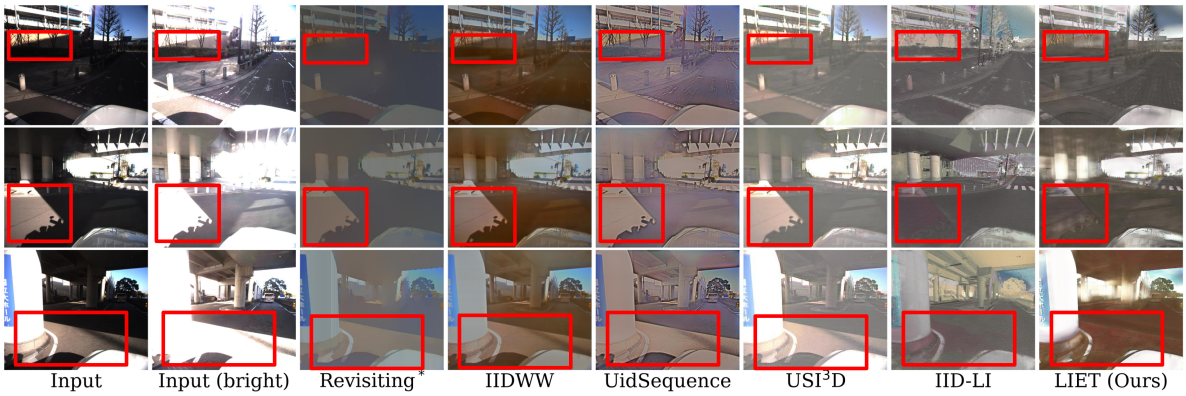


Figure 5. Examples of inferred albedos from various existing models and LIET (ours) with NTT-IID dataset [54]. The compared models include Revisiting* [16], IIDWW [40], UidSequence [36], USI^{3}D [42], and IID-LI [54]. Shadows are less noticeable on the IID-LI and LIET, while cast shadows are visibly retained on the existing models without LiDAR intensity utilization as marked by red square.

same model structures and parameters as USI^{3}D [42][3] and IID-LI. The style encoders E^p_{X}, the content code encoders E^c_{X}, the decoders G_{X}, and the discriminators D_{X} consist of convolutional layers, down-sampling layers, global average pooling layers, dense layers, and residual blocks. Notably, the residual blocks incorporate adaptive instance normalization (AdaIN) [24], and the AdaIN parameters are dynamically determined using a multi-layer perceptron (MLP). To assess images from both global and local perspectives, a multi-scale discriminator [62] is employed in discriminator D_{X}. Given a style code, style mapping functions f_{I} and f_{L} are constructed by MLP. We empirically set the following values for the hyper parameters: $\lambda_{\mathrm{img}} = 100.0, \lambda_{sty} = 10.0, \lambda_{cnt} = 1.0, \lambda_{KLD} = 1.0, \lambda_{VGG} = 1.0, \lambda_{phy} = 10.0$, and $\lambda_{AA} = 100.0$ as detailed in supplementary materials.

[3]LIET is implemented based on USI^{3}D (https://github.com/DreamtaleCore/USI3D.git)

4.2. Intrinsic image decomposition quality

LIET is compared with energy optimization models including baseline-R [4], baseline-S [4], Retinex [19], Color Retinex [19], Bell et al. [4], and Bi et al. [5]. Additionally, IIDWW [40], UidSequence [36], USI^{3}D [42], and IID-LI [54] are implemented as unsupervised learning models. Furthermore, Revisiting* [16] is also evaluated as a supervised model. As shown in Tab. 1, LIET demonstrates a comparable performance to that of IID-LI which utilizes a single image and LiDAR intensity during inference, despite inputting only a single image in LIET. With all annotation, Revisiting* [16] performed better on the two metrics. This model incorporates an albedo flattening module,

Model	MANIQA(↑)	TReS(↑)	MUSIQ(↑)	HYPERIQA(↑)	DBCNN(↑)
Input	0.664	81.1	59.2	0.595	58.1
USI^{3}D [42]	0.488	53.1	40.1	0.298	35.3
IID-LI [54]	0.460	55.5	43.5	0.285	37.5
LIET (ours)	0.570	75.1	56.3	0.414	46.3
Revisiting* [16]	0.395	55.4	44.9	0.389	38.1

Table 2. Numerical comparison in IQA between LIET and the top three models in IID quality including Revisiting* [16], USI^{3}D [42], and IID-LI [54]. Our findings demonstrate that the images inferred by LIET consistently exhibited the highest image quality across all metrics. This comparable performance can be attributed to the absence of image blurring and collapse.

and these metrics are more favorable for inferring flat albedos due to the annotation bias. Additionally, Revisiting* is difficult to train due to the requirement for a large amount of WHDR annotations. Subsequently, the qualitative results of existing models and LIET are illustrated in Fig. 5. Compared to other IID models, both IID-LI and LIET yield inferred albedos with less noticeable shadows due to the LiDAR intensity utilization. Though Revisiting* [16] exhibits a flattened appearance, leading to favorable quantitative outcomes, cast shadows within the images still remain. On the other hand, inferred shades are depicted in supplementary materials. In unsupervised IID models, treating albedo and shade as styles achieves reasonable performance without supervision. However, these are not strictly styles. Thus, incorporating physical constraints from illumination models and geometry components, to bridge the gap caused by treating reflectance and illumination as styles, could enhance inference performance.

4.3. Image quality

Subsequently, we conducted an IQA comparison between LIET and the top three models for their IID quality: Revisiting* [16], USI^{3}D [42], and IID-LI [54]. As shown in Tab. 2, LIET achieves the highest quality across all five evaluation metrics. Revisiting* [16] is trained with relative gray-scaled albedo between nearby points. Thus the model output tends to reduce saturation and leads to reducing IQA ratings. Additionally, both USI^{3}D and IID-LI enhance IID quality through a smoothing process that assumes local albedo flatness. As a result, the images inferred by these models tend to exhibit blurriness, leading to lower IQA ratings. Conversely, in LIET, since fine shadows derived from cast shadows are absent in LiDAR intensity, the albedo inferred from LiDAR intensity has less variation in luminance. LIET achieves albedo local flatness without using smoothing loss due to the alignment of the albedo inferred from LiDAR intensity with the albedo inferred from the image by AA loss. The effect of smooth loss is described in the next section.

Model	F-score(↑)	WHDR(↓)	Precision(↑)	Recall(↑)
Ours	0.607	0.340	0.649	0.601
w/o $\mathcal{L}^{\mathrm{AA}}$	0.437	0.473	0.497	0.476
w/o inst.	0.489	0.447	0.520	0.505
w/o gray	0.601	0.359	0.623	0.596
w/o ILC paths	0.589	0.361	0.641	0.581

Table 3. Effect of AA loss $\mathcal{L}^{\mathrm{AA}}$. "w/o inst." describes the loss $\mathcal{L}^{\mathrm{AA}}$ calculated without instance normalization in AA loss. "w/o gray" refers to delete the gray scaling from AA loss.

4.4. Ablation study

This section describes an ablation study for the contribution of AA loss and ILC paths. Tab. 3 demonstrates the effect of AA loss due to the direct connection between the image-encoder path and the LiDAR-encoder path during training. Since the distribution of LiDAR intensity varies across samples, features are well-trained by applying instance normalization rather than by scaling uniformly across all samples. The inference quality is slightly improved by aligning these albedos in gray scale, due to the lack of hue in LiDAR intensity. Additionally, the ILC paths contributes separating an image into content and style codes by mutually translating the image and LiDAR intensity, which share the contents but differ styles. Thus, the IID quality is improved by ILC paths. Without the stop gradient in $\mathcal{L}^{\mathrm{AA}}$, X_{RI} and X_{RL} tend to converge to flat images to simply minimize $\mathcal{L}^{\mathrm{AA}}$, resulting in unstable training and undesirable outputs. The ablation study of the model architecture is described in supplementary material.

5. Conclusion

In this paper, we proposed *unsupervised single-image intrinsic image decomposition with LiDAR intensity enhanced training (LIET)*. We proposed a novel approach in which an image and LiDAR intensity are individually fed into the model during training, while the inference process only employs a single image. To calculate the relationship between the image-encoder path and the LiDAR-encoder path, we introduced AA loss to align the albedo inferred from a single image to that from LiDAR intensity, and ILC paths to enhance the separation of contents and styles. As a result, LIET achieved performance comparable to state-of-the-art in IID quality metrics while only employing a single image as input during inference. Furthermore, LIET demonstrated improvements in image quality supported by the five most recent IQA metrics. While this study focused on Lambertian surfaces for IID, utilizing additional domain data (hyperspectral image/ multi-bounce LiDAR) could extend the applicability of LIET to diffuse-specular mixed surfaces [15, 22].

References

[1] Harry Barrow, J Tenenbaum, A Hanson, and E Riseman. Recovering Intrinsic Scene Characteristics from Images. *Computer Vision Systems*, 2(3-26):2, 1978. 1

[2] Anil S Baslamisli, Thomas T Groenestege, Partha Das, Hoang-An Le, Sezer Karaoglu, and Theo Gevers. Joint Learning of Intrinsic Images and Semantic Segmentation. In *ECCV*, pages 286–302, 2018. 1

[3] Shida Beigpour and Joost Van De Weijer. Object recoloring based on intrinsic image estimation. In *ICCV*, pages 327–334, 2011. 1

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

[5] 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, 7

[6] Mark Boss, Raphael Braun, Varun Jampani, Jonathan T Barron, Ce Liu, and Hendrik Lensch. NeRD: Neural Reflectance Decomposition from Image Collections. In *ICCV*, pages 12684–12694, 2021. 3

[7] Mark Boss, Varun Jampani, Raphael Braun, Ce Liu, Jonathan Barron, and Hendrik Lensch. Neural-PIL: Neural Pre-Integrated Lighting for Reflectance Decomposition. *NeurIPS*, 34:10691–10704, 2021. 3

[8] 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

[9] 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, 2012. 2

[10] 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. 2

[11] Qifeng Chen and Vladlen Koltun. A Simple Model for Intrinsic Image Decomposition with Depth Cues. In *ICCV*, pages 241–248, 2013. 1

[12] Qifeng Chen and Vladlen Koltun. Photographic Image Synthesis with Cascaded Refinement Networks. In *ICCV*, pages 1511–1520, 2017. 6

[13] Bin Cheng, Zuhao Liu, Yunbo Peng, and Yue Lin. General Image-to-Image Translation with One-Shot Image Guidance. In *ICCV*, pages 22736–22746, 2023. 3

[14] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation. In *CVPR*, pages 8789–8797, 2018. 2

[15] Gamal ElMasry, Pere Gou, and Salim Al-Rejaie. Effectiveness of specularity removal from hyperspectral images on the quality of spectral signatures of food products. *Journal of Food Engineering*, 289:110148, 2021. 8

[16] Qingnan Fan, Jiaolong Yang, Gang Hua, Baoquan Chen, and David Wipf. Revisiting Deep Intrinsic Image Decompositions. In *CVPR*, pages 8944–8952, 2018. 1, 2, 7, 8

[17] S Alireza Golestaneh, Saba Dadsetan, and Kris M Kitani. No-Reference Image Quality Assessment via Transformers, Relative Ranking, and Self-Consistency. In *WACV*, pages 1220–1230, 2022. 2, 6

[18] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. *NeurIPS*, 27:139–144, 2014. 2, 5

[19] 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, 2009. 1, 2, 7

[20] Maximilien Guislain, Julie Digne, Raphaëlle Chaine, Dimitri Kudelski, and Pascal Lefebvre-Albaret. Detecting and Correcting Shadows in Urban Point Clouds and Image Collections. In *3DV*, pages 537–545, 2016. 3

[21] Jon Hasselgren, Nikolai Hofmann, and Jacob Munkberg. Shape, Light, and Material Decomposition from Images using Monte Carlo Rendering and Denoising. *NeurIPS*, 35:22856–22869, 2022. 3

[22] Connor Henley, Siddharth Somasundaram, Joseph Hollmann, and Ramesh Raskar. Detection and mapping of specular surfaces using multibounce lidar returns. *Optics Express*, 31(4):6370–6388, 2023. 8

[23] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. *NeurIPS*, 33:6840–6851, 2020. 3

[24] Xun Huang and Serge Belongie. Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization. In *ICCV*, pages 1501–1510, 2017. 7

[25] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal Unsupervised Image-to-Image Translation. In *ECCV*, pages 172–189, 2018. 3

[26] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-Image Translation with Conditional Adversarial Networks. In *CVPR*, pages 1125–1134, 2017. 2

[27] Junho Jeon, Sunghyun Cho, Xin Tong, and Seungyong Lee. Intrinsic Image Decomposition Using Structure-Texture Separation and Surface Normals. In *ECCV*, pages 218–233, 2014. 1

[28] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual Losses for Real-Time Style Transfer and Super-Resolution. In *ECCV*, pages 694–711, 2016. 6

[29] 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

[30] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. MUSIQ: Multi-scale Image Quality Transformer. In *ICCV*, pages 5148–5157, 2021. 2, 6

[31] Philipp Krähenbühl. Free Supervision From Video Games. In *CVPR*, pages 2955–2964, 2018. 1, 6

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

[33] 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

[34] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse Image-to-Image Translation via Disentangled Representations. In *ECCV*, pages 35–51, 2018. 3

[35] 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, 2012. 1

[36] Louis Lettry, Kenneth Vanhoey, and Luc Van Gool. Unsupervised Deep Single Image Intrinsic Decomposition using Illumination Varying Image Sequences. *Computer Graphics Forum*, 37, 2018. 2, 7

[37] Bo Li, Kaitao Xue, Bin Liu, and Yu-Kun Lai. BBDM: Image-to-image Translation with Brownian Bridge Diffusion Models. In *CVPR*, pages 1952–1961, 2023. 3

[38] Enxu Li, Sergio Casas, and Raquel Urtasun. MemorySeg: Online LiDAR Semantic Segmentation with a Latent Memory. In *ICCV*, pages 745–754, 2023. 3

[39] Zhengqi Li and Noah Snavely. CGIntrinsics: Better Intrinsic Image Decomposition through Physically-Based Rendering. In *ECCV*, pages 371–387, 2018. 2

[40] Zhengqi Li and Noah Snavely. Learning Intrinsic Image Decomposition from Watching the World. In *CVPR*, pages 9039–9048, 2018. 2, 7

[41] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised Image-to-Image Translation Networks. *NeurIPS*, 30, 2017. 2

[42] 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. 1, 2, 3, 6, 7, 8

[43] 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

[44] 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, 3

[45] 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

[46] Abhimitra Meka, Michael Zollhöfer, Christian Richardt, and Christian Theobalt. Live Intrinsic Video. *ACM TOG*, 35(4):1–14, 2016. 1

[47] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 3

[48] Jacob Munkberg, Jon Hasselgren, Tianchang Shen, Jun Gao, Wenzheng Chen, Alex Evans, Thomas Müller, and Sanja Fidler. Extracting Triangular 3D Models, Materials, and Lighting From Images. In *CVPR*, pages 8280–8290, 2022. 3

[49] Takuya Narihira, Michael Maire, and Stella X Yu. Direct Intrinsics: Learning Albedo-Shading Decomposition by Convolutional Regression. In *ICCV*, pages 2992–2992, 2015. 1

[50] Takuya Narihira, Michael Maire, and Stella X Yu. Learning Lightness From Human Judgement on Relative Reflectance. In *CVPR*, pages 2965–2973, 2015. 1

[51] 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

[52] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image Super-Resolution via Iterative Refinement. *IEEE TPAMI*, 45(4):4713–4726, 2022. 3

[53] Shogo Sato, Yasuhiro Yao, Taiga Yoshida, Shingo Ando, and Jun Shimamura. Shadow Detection Based on Luminance-LiDAR Intensity Uncorrelation. *IEICE Transactions*, 106(9):1556–1563, 2023. 3

[54] Shogo Sato, Yasuhiro Yao, Taiga Yoshida, Takuhiro Kaneko, Shingo Ando, and Jun Shimamura. Unsupervised Intrinsic Image Decomposition With LiDAR Intensity. In *CVPR*, pages 13466–13475, 2023. 1, 2, 3, 6, 7, 8

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

[56] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *ICLR*, pages 1–14, 2015. 6

[57] Pratul P Srinivasan, Boyang Deng, Xiuming Zhang, Matthew Tancik, Ben Mildenhall, and Jonathan T Barron. NeRV: Neural Reflectance and Visibility Fields for Relighting and View Synthesis. In *CVPR*, pages 7495–7504, 2021. 3

[58] Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqiu Sun, and Yanning Zhang. Blindly Assess Image Quality in the Wild Guided by a Self-Adaptive Hyper Networkk. In *CVPR*, pages 3667–3676, 2020. 2, 6

[59] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance Normalization: The Missing Ingredient for Fast Stylization. *arXiv preprint arXiv:1607.08022*, 2016. 6

[60] Ben Upcroft, Colin McManus, Winston Churchill, Will Maddern, and Paul Newman. Lighting Invariant Urban Street Classification. In *ICRA*, pages 1712–1718, 2014. 1

[61] Chenglin Wang, Yunchao Tang, Xiangjun Zou, Weiming SiTu, and Wenxian Feng. A robust fruit image segmentation algorithm against varying illumination for vision system of fruit harvesting robot. *Optik*, 131:626–631, 2017. 1

[62] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs. In *CVPR*, pages 8798–8807, 2018. 6, 7

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

[64] Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. MANIQA: Multi-dimension Attention Network for No-Reference Image Quality Assessment. In *CVPRW*, pages 1191–1200, 2022. 2, 6

[65] Ye Yu and William AP Smith. InverseRenderNet: Learning single image inverse rendering. In *CVPR*, pages 3155–3164, 2019. 3

[66] Kai Zhang, Fujun Luan, Qianqian Wang, Kavita Bala, and Noah Snavely. PhySG: Inverse Rendering with Spherical Gaussians for Physics-based Material Editing and Relighting. In *CVPR*, pages 5453–5462, 2021. 3

[67] Weixia Zhang, Kede Ma, Jia Yan, Dexiang Deng, and Zhou Wang. Blind Image Quality Assessment Using A Deep Bilinear Convolutional Neural Network. *IEEE TCSVT*, 30(1):36–47, 2018. 2, 6

[68] Xiuming Zhang, Pratul P Srinivasan, Boyang Deng, Paul Debevec, William T Freeman, and Jonathan T Barron. NeRFactor: Neural Factorization of Shape and Reflectance Under an Unknown Illumination. *ACM TOG*, 40(6):1–18, 2021. 3

[69] 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

[70] Hao Zhou, Xiang Yu, and David W Jacobs. GLoSH: Global-Local Spherical Harmonics for Intrinsic Image Decomposition. In *ICCV*, pages 7820–7829, 2019. 1

[71] Tinghui Zhou, Philipp Krahenbuhl, and Alexei A Efros. Learning Data-driven Reflectance Priors for Intrinsic Image Decomposition. In *ICCV*, pages 3469–3477, 2015. 1

[72] Jingsen Zhu, Yuchi Huo, Qi Ye, Fujun Luan, Jifan Li, Dianbing Xi, Lisha Wang, Rui Tang, Wei Hua, Hujun Bao, et al. I2-SDF: Intrinsic Indoor Scene Reconstruction and Editing via Raytracing in Neural SDFs. In *CVPR*, pages 12489–12498, 2023. 3

[73] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. In *ICCV*, pages 2223–2232, 2017. 2

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