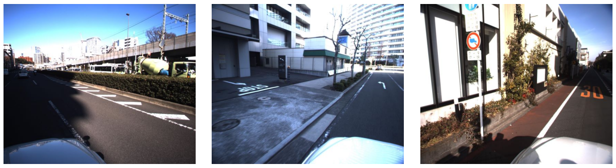
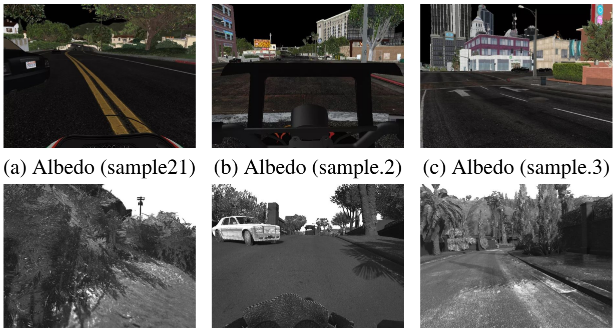
Unsupervised Intrinsic Image Decomposition with LiDAR Intensity Supplementary Material

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



(a) Shinagawa (b) Yokohama (c) Mitaka

Figure 1. Examples of observed outdoor scenes measured at Shinagawa, Yokohama and Mitaka in Japan.



(a) Albedo (sample21) (b) Albedo (sample.2) (c) Albedo (sample.3)

(d) Shade (sample.1) (e) Shade (sample.2) (f) Shade (sample.3)

Figure 2. Examples of data for albedo and shade domains from an FSVG dataset [6].

1. Prepared datasets

In this paper, we prepared two types of datasets. One is a dataset observed with a LiDAR and a RGB camera in real scenes. The other is a dataset of albedos and shades from the free supervision from video games (FSVG) dataset [6].

For the observed dataset, outdoor scenes were measured with a camera and LiDAR at Shinagawa, Yokohama and Mitaka in Japan. As shown in Fig. 1, the data for these regions have different characteristics. First, Shinagawa was measured on a wide street, thus showing many roadways and cars. In Yokohama, measurements were taken on a relatively narrow roadway in a group of buildings. Furthermore, most of the data in Mitaka were measured on a very narrow road between houses. By combining data from different measurement points in this way, we were able to increase the variation of the data. Our own dataset is currently being prepared for release.

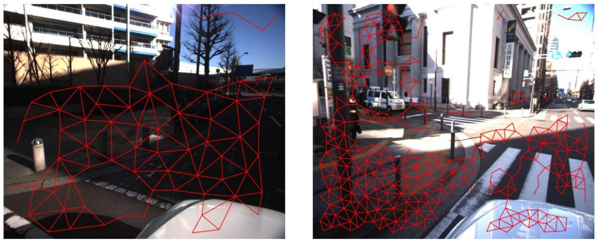


Figure 3. Examples of sparse annotation (left) and dense annotation (right). The points to be compared are connected by red lines.

For the albedo and shade domain datasets, we used FSVG datasets. FSVG is a synthetic image dataset of outdoor scenes, for which a set of albedos and rendered images have been prepared. Since there was no shade image, shades were generated for dividing the rendered image by albedo based on Retinex theory [7] in Eq. (1).

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

In order to perform inference on the city data, we selected scenes with buildings as much as possible. Examples of albedo and shade used for training are shown in Fig. 2. Since the proposed method in this study is unsupervised learning, albedo and shade were extracted from 10000 samples of completely independent data.

2. Annotation

For quantitative evaluation for IID, we annotated the relative reflectance intensity in the same manner as Bell et al [1]. First, we extracted 100 samples and 10 samples from the obtained dataset for sparse and dense annotation, re-

λ_7	WHDR	precision	recall	F-score
0.0	0.455	0.513	0.473	0.430
1.0	0.440	0.543	0.491	0.453
3.0	0.389	0.604	0.545	0.551
5.0	0.353	0.627	0.587	0.596
10.0	0.356	0.630	0.580	0.589
20.0	0.353	0.625	0.596	0.602

Table 1. Ablation study for the weight of intensity consistency loss λ_7 with our dataset for randomly sampled annotation points.

spectively. Each example for annotation is shown in Fig. 3. We labeled each edge with relative intensity. Although, points where reflectance could not be defined such as sky or windows were labeled "NG" and excluded from the evaluation. In addition, points around a strong edge, saturated regions and significant variances were removed, hence, evaluation points are concentrated in the low-frequency region. Thus, the amount of annotations are considered to be biased ("E" >> "L" or "D"), which is the reason why we evaluated random sampled data.

3. Ablation study for intensity consistency loss

In original USI^{3}D [10], a weighted sum of the seven losses is optimized.

$$\begin{aligned}\min_{E,G,f}\max_{D}(E,G,f,D) = \mathcal{L}^{\text{adv}} + \lambda_1\mathcal{L}^{\text{cnt}} + \lambda_2\mathcal{L}^{\text{KL}} \\ + \lambda_3\mathcal{L}^{\text{img}} + \lambda_4\mathcal{L}^{\text{pri}} + \lambda_5\mathcal{L}^{\text{phy}} + \lambda_6\mathcal{L}^{\text{smooth}},\end{aligned} \quad (2)$$

In the original paper, λ_1, λ_2, λ_3, λ_4, λ_5, and λ_6 are set as 10.0, 0.1, 10.0, 0.1, 5.0, and 1.0, respectively. In this paper, we designed an intensity consistency loss that aims to provide a criterion for the ill-posed problem of decomposing a single image. Thus, we added $\mathcal{L}^{\text{int}}$ with weight parameter λ_7 to Eq. (2). To obtain a better result, we searched weight λ_7 as shown in Tab. 1. Although there was no significant difference in the range of 5 to 20, the highest F-Score ($\lambda_7 = 20$) was used in this study.

4. LiDAR intensity densification

A LiDAR intensity densification (LID) module is used for robustness for LiDAR sparsity or occlusions. In this section, the effect of the LID module is compared with that of the conventional methods and the original deep image prior (DIP) [12]. As a conventional method, a Navier-Stokes (NS) based algorithm [2] and a fast-marching method (FMM) based algorithm [11] are implemented. In Fig. 4, "ground truth" represents the observed-LiDAR intensity, and "input" is LiDAR intensity with 1% density of the observed data. Since the conventional methods completed LiDAR intensity without the RGB image, the generated image was blurred in sparse regions. On the other hand, the LID module received the RGB image, hence LiDAR intensity is densified while considering image edges and brightness. These characteristics are noticeable in the building and white line areas in Fig. 4. Thus, we selected a LID module for LiDAR intensity densification.

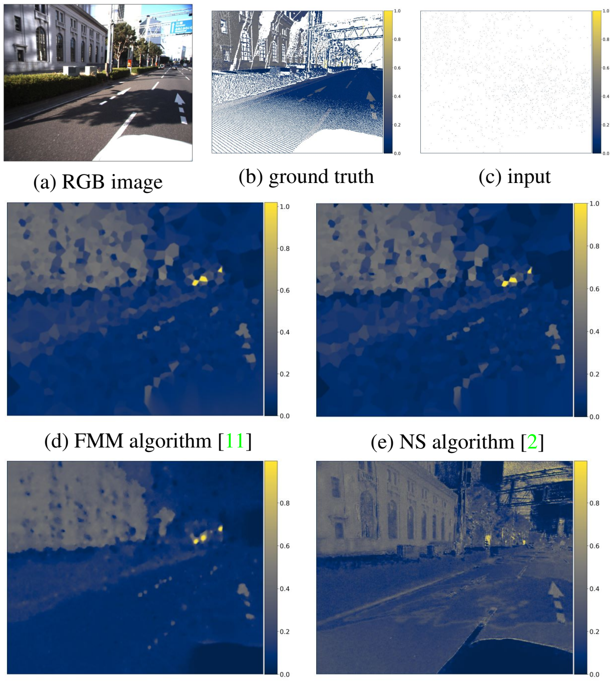

(a) RGB image
(b) ground truth
(c) input
(d) FMM algorithm [11]
(e) NS algorithm [2]
(f) DIP [12]
(g) LID module

Figure 4. Comparing the effect of an LID module with that of the conventional methods [2, 11] and original DIP [12]. Points with no values are shown in white.

λ_8	WHDR	precision	recall	F-score
0.0	0.353	0.625	0.596	0.602
1.0	0.355	0.620	0.593	0.600
5.0	0.367	0.602	0.593	0.596
10.0	0.362	0.608	0.592	0.597
20.0	0.361	0.609	0.599	0.602

Table 2. Ablation study for the weight of hue consistency loss λ_8 with our dataset for randomly sampled annotation points.

5. Hue consistency loss

In both the conventional methods [1, 3–5, 8–10] and IID-LI, the estimated albedo hue may deviate significantly from the input image, since the irradiation light is sometimes colorful, such as the setting sun. When the irradiated light can be approximated as white light, the estimated albedo and the input image are considered to have similar hue. Thus, in

methods	WHDR	precision	recall	F-score
USI^{3}D	0.326±0.024	0.433±0.007	0.502±0.003	0.425±0.012
Ours (without LID)	0.297±0.015	0.451±0.006	0.522±0.005	0.457±0.008
Ours (without $\mathcal{L}^{\text{int}}$)	0.364±0.065	0.417±0.012	0.466±0.011	0.399±0.024
Ours (without gamma correction)	0.439±0.012	0.442±0.002	0.563±0.003	0.412±0.005
Ours	0.236±0.014	0.507±0.010	0.582±0.009	0.514±0.010

Table 3. The average estimation quality of the five trials in our dataset for all (E=9411, D=2554, L=661) annotation points.

methods	WHDR	precision	recall	F-score
USI^{3}D	0.430±0.006	0.511±0.024	0.499±0.002	0.444±0.006
Ours (without LID)	0.420±0.009	0.534±0.009	0.522±0.006	0.524±0.006
Ours (without $\mathcal{L}^{\text{int}}$)	0.471±0.016	0.475±0.030	0.461±0.008	0.416±0.008
Ours (without gamma correction)	0.434±0.006	0.555±0.006	0.563±0.006	0.551±0.006
Ours	0.363±0.008	0.613±0.010	0.584±0.010	0.590±0.010

Table 4. The average estimation quality of the five trials in our dataset for randomly sampled (E=661, D=661, L=661) annotation points.

such a case, inconsistency in hue can be reduced by adding hue consistency loss in Eq. (3).

$$\mathcal{L}^{\text{hue}} = |H(R(I)) - H(I)|, \tag{3}$$

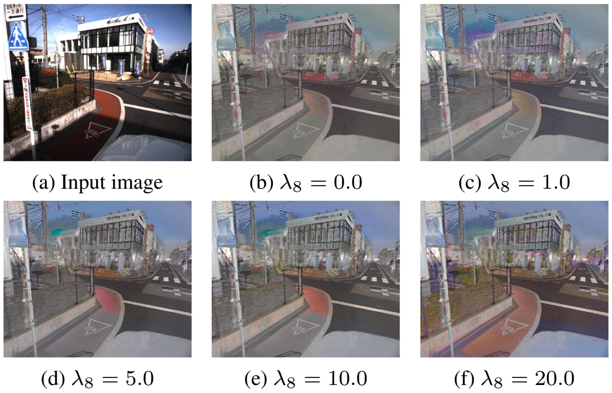
where $H(x)$ is a function that extract the hue of image x. In summary, the $\mathcal{L}^{\text{hue}}$ is added to Eq. (2) with the weight parameter λ_8.

$$\min_{E,G,f} \max_{D}(E, G, f, D) = \mathcal{L}^{\text{adv}} + \lambda_1\mathcal{L}^{\text{cnt}} + \lambda_2\mathcal{L}^{\text{KL}} \\ +\lambda_3\mathcal{L}^{\text{img}}+\lambda_4\mathcal{L}^{\text{pri}}+\lambda_5\mathcal{L}^{\text{phy}}+\lambda_6\mathcal{L}^{\text{smooth}}+\lambda_7\mathcal{L}^{\text{int}}+\lambda_8\mathcal{L}^{\text{hue}}, \tag{4}$$

To obtain a better result, we searched weight λ_8 as shown in Tab. 2. Since hue is not relevant in the evaluation of this study, the estimation accuracy was approximately constant. Fig. 5 shows the qualitative evaluation results. As λ_8 was increased, the hue of the input and the estimation got closer. In particular, this loss was effective for the sidewalk with red in Fig. 5. When irradiated light can be approximated as white, hue consistency loss is considered to be effective for visual improvement.

6. Variation in estimation quality

In this paper, we evaluated the estimation quality of conventional methods [1, 3–5, 8–10] and IID-LI with our dataset. For training USI^{3}D [10] and IID-LI, we performed five trials per condition due to initial value dependence, and listed the best performance in the table in the main manuscript. Since the variation in estimation quality was not obtained from the listing, the mean values and standard deviations for each experiment are shown in Tabs. 3 and 4. As a whole, the variation in the accuracy of IID-LI was not much different from that of USI^{3}D. Tabs. 3 and 4 show that IID-LI is superior to USI^{3}D in accuracy, even accounting for errors.


(a) Input image (b) $\lambda_8 = 0.0$ (c) $\lambda_8 = 1.0$

(d) $\lambda_8 = 5.0$ (e) $\lambda_8 = 10.0$ (f) $\lambda_8 = 20.0$

Figure 5. Estimating examples for IID-LI with hue consistency loss. As λ_8 was increased, the hue of the input and the estimation got closer.

References

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

[2] Marcelo Bertalmio, Andrea L Bertozzi, and Guillermo Sapiro. Navier-stokes, fluid dynamics, and image and video inpainting. In *CVPR*, volume 1, pages I–I. IEEE, 2001. 2

[3] 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. 2, 3

[4] Qingnan Fan, Jiaolong Yang, Gang Hua, Baoquan Chen, and David Wipf. Revisiting deep intrinsic image decompositions. In *CVPR*, pages 8944–8952, 2018. 2, 3

[5] 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. 2, 3

[6] Philipp Krähenbühl. Free supervision from video games. In *CVPR*, pages 2955–2964, 2018. 1

[7] 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

[8] Louis Lettry, Kenneth Vanhoey, and Luc Van Gool. Unsupervised Deep Single-Image Intrinsic Decomposition using Illumination-Varying Image Sequences. *Comput. Graph. Forum*, 37, 2018. 2, 3

[9] Zhengqi Li and Noah Snavely. Learning intrinsic image decomposition from watching the world. In *CVPR*, pages 9039–9048, 2018. 2, 3

[10] 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

[11] Alexandru Telea. An image inpainting technique based on the fast marching method. *Journal of graphics tools*, 9(1):23–34, 2004. 2

[12] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *CVPR*, pages 9446–9454, 2018. 2