Hierarchical Disentangled Representation Learning for Outdoor Illumination Estimation and Editing

Piaopiao Yu[1], Jie Guo[1,†], Fan Huang[1], Cheng Zhou[1], Hongwei Che[2], Xiao Ling[2], Yanwen Guo[1,†]
[1]National Key Lab for Novel Software Technology, Nanjing University, China
[2]Guangdong OPPO Mobile Telecommunications Corp Ltd, China
turmikey@163.com, {guojie, ywguo}@nju.edu.cn
{mf20330031, zc}@smail.nju.edu.cn, {chehongwei, lingxiao}@oppo.com

Abstract

*Data-driven sky models have gained much attention in outdoor illumination prediction recently, showing superior performance against analytical models. However, naively compressing an outdoor panorama into a low-dimensional latent vector, as existing models have done, causes two major problems. One is the mutual interference between the HDR intensity of the sun and the complex textures of the surrounding sky, and the other is the lack of fine-grained control over independent lighting factors due to the entangled representation. To address these issues, we propose a hierarchical disentangled sky model (HDSky) for outdoor illumination prediction. With this model, any outdoor panorama can be hierarchically disentangled into several factors based on three well-designed autoencoders. The first autoencoder compresses each sunny panorama into a sky vector and a sun vector with some constraints. The second autoencoder and the third autoencoder further disentangle the sun intensity and the sky intensity from the sun vector and the sky vector with several customized loss functions respectively. Moreover, a unified framework is designed to predict all-weather sky information from a single outdoor image. Through extensive experiments, we demonstrate that the proposed model significantly improves the accuracy of outdoor illumination prediction. It also allows users to intuitively edit the predicted panorama (*e.g.*, changing the position of the sun while preserving others), without sacrificing physical plausibility.*

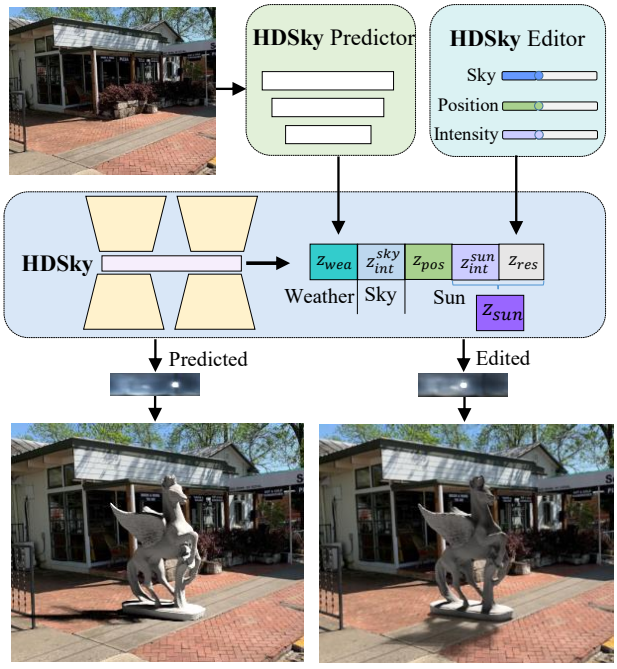

Figure 1. We propose HDSky, the first disentangled data-driven model for all-weather outdoor illumination. As the learned latent vector is disentangled into several independent and meaningful factors, state-of-the-art performance on outdoor illumination prediction can be achieved, leading to consistent shading and shadows in AR rendering and enabling intuitive illumination editing with physical plausibility.

1. Introduction

Outdoor illumination prediction based on a single input image is a key task for many applications ranging from scene understanding and reconstruction to augmented reality (AR). However, the diverse weather conditions and the complex interaction between illumination and other scene properties (*e.g.*, surface reflectance and geometric variations) make this problem highly challenging.

Due to the success of deep learning, learning-based methods [9, 4, 33, 8] have emerged recently and achieved state-of-the-art performance by employing powerful deep neural networks to automatically learn the mapping between an image of limited field-of-view (FoV) and a given

†Corresponding authors.

sky model. To make the problem tractable, some approaches resort to analytical sky models. For instance, Hold-Geoffroy *et al*. [9] adopted the Hošek-Wilkie sky (HW) model [10, 23] to encode 360° high dynamic range (HDR) illumination with as few as 4 parameters. However, this model only works well for clear skies. Zhang *et al*. [33] employed the Lalonde-Matthews (LM) model [17] to represent all-weather outdoor illumination, but they cannot faithfully predict the sky color due to the limited expressiveness of the LM model. Generally, analytical models fail to fully capture the complexities of real-world atmospheric conditions.

Recently, data-driven sky models [28, 4, 8], which compress complex outdoor illumination into a low-dimensional latent vector, become popular in outdoor illumination prediction and achieve state-of-the-art performance. After training on large-scale datasets, data-driven sky models can reproduce a much wider range of outdoor illumination than most analytical models, with far less bias. However, naively compressing an outdoor panorama into a low-dimensional latent vector with neural networks would cause two major problems. First, a single vector is hard to reflect the HDR intensity of the sun and the diverse weather conditions simultaneously. Consequently, the intensity of the sun would be lowered in order to match the correct sky color and textures, leading to low accuracy in prediction and inconsistent shading and shadows in AR applications. Second, intuitively editing outdoor illumination based on a single latent vector is challenging due to the conflated representation. For instance, it is difficult to change the position of the sun while preserving other properties.

To tackle the above problems, we propose HDSky, a new data-driven sky model which is hierarchically trained on HDR panoramas to disentangle each outdoor HDR panorama into several meaningful factors. After manually classifying an input panorama into sunny or cloudy, three autoencoders are well designed to generate the disentangled representation for sunny panoramas. The first autoencoder compresses each sunny panorama into a sky vector and a sun vector by imposing constraints on the sky and sun based on the information theory. These two vectors are expected to recover the original panorama after proper fusion. The second autoencoder further disentangles the sun vector into the sun intensity and the residual factor with several customized loss functions. The third autoencoder disentangles the sky intensity from the sky vector. For the cloudy panorama, a single autoencoder is required to compress the panorama into a sky vector. Another autoencoder is also utilized to disentangle the sky intensity from the sky vector.

With HDSky, we are able to achieve higher accuracy in outdoor illumination prediction than previous methods based on conflated representations [8], since the entanglement among different sky properties (*e.g.*, the intensity of the sun and the color of the sky) is avoided. Specifically, for an FoV-limited image, we first predict its weather condition (sunny or cloudy) and then estimate each individual factor using several CNNs which are trained jointly on the SUN360 dataset [32] with properly designed loss functions. Utilizing the trained decoders of HDSky, we can fuse the disentangled factors and recover the full HDR panorama that can be used directly in AR rendering, guaranteeing consistent shading and shadows. The disentangled representation allows us to intuitively edit the predicted outdoor illumination, achieving more vivid results with limited manual assistance.

The contributions of our work can be summarized as follows.

- We introduce HDSky, a novel data-driven sky model that hierarchically disentangles an outdoor panorama into several interpretable vectors based on the information theory.
- We propose a unified framework that can predict all-weather sky information from a single outdoor image, achieving state-of-the-art performance in outdoor illumination prediction.
- We develop an intuitive editing tool based on HDSky that allows to alter the predicted outdoor illumination with fine-grained control and physical plausibility.

2. Related work

Traditional methods for outdoor lighting estimation. Based on the Perez model [25], Lalonde *et al*. [16] utilized hand-crafted priors (*e.g.*, cast shadows) to recover lighting from a single, generic outdoor image. Karsch *et al*. [13] can extrapolate the scene outside the field of view and estimate the out-of-view illumination by matching the input image to the SUN360 panorama [32], which can not be directly linked with illumination. The linearity of light transport is leveraged in several works to estimate lighting from faces [29, 30]. Some works on illumination estimation assumed that the geometry is known and relied on strong priors on scene reflectance, geometry and illumination [3, 2, 22]. Typically, these approaches do not generalize to large-scale outdoor scenes.

Deep learning for outdoor lighting estimation. Recently, significant progress has been made in lighting estimation with deep learning methods. For example, a reflectance map was inferred from a single image of an object with known geometry [27] and can be further factored into lighting and material properties [7]. Given two opposing views of a panorama, Cheng *et al*. [6] estimated lighting using a deep learning technique. Hold-Geoffroy *et al*. [9] utilized the parametric Hošek-Wilkie sky model [10, 23] to model outdoor lighting and learn to estimate its parameters

from a single image. Subsequently, Zhang *et al*. [33] extended this method with the Lalonde-Matthews outdoor illumination model. However, for these analytical sky models [25, 26, 10, 23, 17], it is challenging to accurately represent the complex weather conditions with only a few parameters. In contrast, we estimate more plausible and accurate lighting in a data-driven manner.

Liu *et al*. [21] generate plausible virtual object shadows with neural networks. Several deep learning techniques for inverse lighting rely on cues from faces. Zhou *et al*. [34] estimated a low-frequency 2nd order spherical harmonic illumination from portraits. For higher frequency lighting estimates, millions of LDR images of three diffuse, glossy, and mirror reference spheres are used to train a model to regress the omnidirectional HDR lighting from an indoor or outdoor image [18]. Subsequently, LeGendre *et al*. [19] extended previous methods with diverse portraits and achieved superior performance. Calian *et al*. [4] employed a deep autoencoder to learn a data-driven model and estimated the HDR outdoor lighting from a single face image. Similarly, Hold-Geoffroy *et al*. [8] utilized an autoencoder to estimate lighting from a single image of a generic outdoor scene in an end-to-end framework. Unfortunately, the learned latent vector of the representation is not interpretable, which limits its accuracy and applicability.

Disentangled representation learning. Disentanglement learning aims to disentangle the underlying factors that form real-world data [5, 12]. While many supervised methods require strong supervision (edge/keypoint/mask annotations or detectors) [24, 1], most unsupervised methods are limited to disentangling at most two factors like shape and texture. In this paper, we factor the outdoor HDR panorama into more factors such as the sky, the sun position, and the sun intensity. Several methods leverage information theory [5] to disentangle the underlying factors with minimal supervision [31, 20]. We adopt this theory in our framework to disentangle outdoor illumination.

3. Overview

Our goal in this paper is to predict outdoor illumination from a single FoV-limited image and to allow convenient editing of the predicted illumination for further user control (*e.g.*, fine-tune the sun intensity). To achieve this goal and ensure high accuracy, we resort to a data-driven, hierarchical disentangled sky model *HDSky*, which is learned from outdoor HDR panoramas.

Before training HDSky, we first manually divide all outdoor HDR panoramas into two categories: sunny ($z_{wea} = 1$) and cloudy ($z_{wea} = 0$). For the sunny panorama, we utilize three well-designed autoencoders to hierarchically learn the disentangled representation. The first autoencoder AE_1 factors each sunny panorama $\mathcal{P}$ into a sky vector z_{sky} and a sun vector z_{sun} according to the information theory. The full panorama $\mathcal{P}'$ then can be recovered from z_{sky}, z_{sun} and the ground-truth sun position z_{pos} with two decoders. The second autoencoder AE_2 and the third autoencoder AE_3 further disentangle the sun intensity z_{int}^{sun} and the sky intensity z_{int}^{sky} from the sun vector z_{sun} and the sky vector z_{sky} respectively. The cloudy panorama is compressed by a single autoencoder into a sky vector. We then employ another autoencoder to learn the sky intensity from the sky vector of the cloudy panorama. The disentangled factors of the illumination latent space and the corresponding dimension are shown in the supplementary material.

To perform single-image outdoor illumination prediction based on HDSky, we first classify the input image into either sunny or cloudy with a network. Whereafter, different neural networks are used to estimate the disentangled vectors. Finally, we can recover the full HDR panorama $\hat{\mathcal{P}}$ by fusing the predicted vectors with the trained decoders of HDSky.

The disentanglement representation also allows us to edit the predicted outdoor illumination more conveniently. We can edit the sun intensity, the sun position and the sky intensity intuitively and independently.

4. HDSky

In this section, we describe our HDSky and the training process in detail.

4.1. Network architecture

Our HDSky handles sunny and cloudy panoramas differently. With three autoencoders AE_1, AE_2 and AE_3, HDSky hierarchically compresses each sunny HDR panorama $\mathcal{P}$ into a low-dimensional latent vector with several disentangled factors, as illustrated in the left part of Fig. 2.

The first autoencoder AE_1 disentangles $\mathcal{P}$ into a sky vector z_{sky}, a sun vector z_{sun} with two parallel encoders. As shown in the top-left corner of Fig. 2, the full panorama $\mathcal{P}'$ can be reconstructed by fusing the above two vectors with the ground-truth sun position vector z_{pos} using the well-designed decoders. Two panoramas $\mathcal{P}_{sky}$ and $\mathcal{P}_{sun}$ are generated during the reconstruction. The sun panorama $\mathcal{P}_{sun}$ is generated by the sun decoder which is conditioned on the sun position map $\mathcal{P}_{pos}$ produced by z_{pos}. In our current implementation, $\mathcal{P}_{pos}$ is a binary image indicating the position of the sun in the panorama. Then, the full panorama $\mathcal{P}'$ is reconstructed by stitching the sun panorama on the sky panorama with the sun position map, *i.e.*,

$$\mathcal{P}' = \mathcal{P}_{sun} + \mathcal{P}_{sky} \odot (1 - \mathcal{P}_{pos}), \tag{1}$$

in which $\odot$ is element-wise multiplication. To achieve disentanglement for AE_1, we use the information theory to enforce high mutual information between (1) z_{sky} and $\mathcal{P}_{sky}$, and (2) z_{sun} and $\mathcal{P}_{sun}$. This is detailed in the next section.

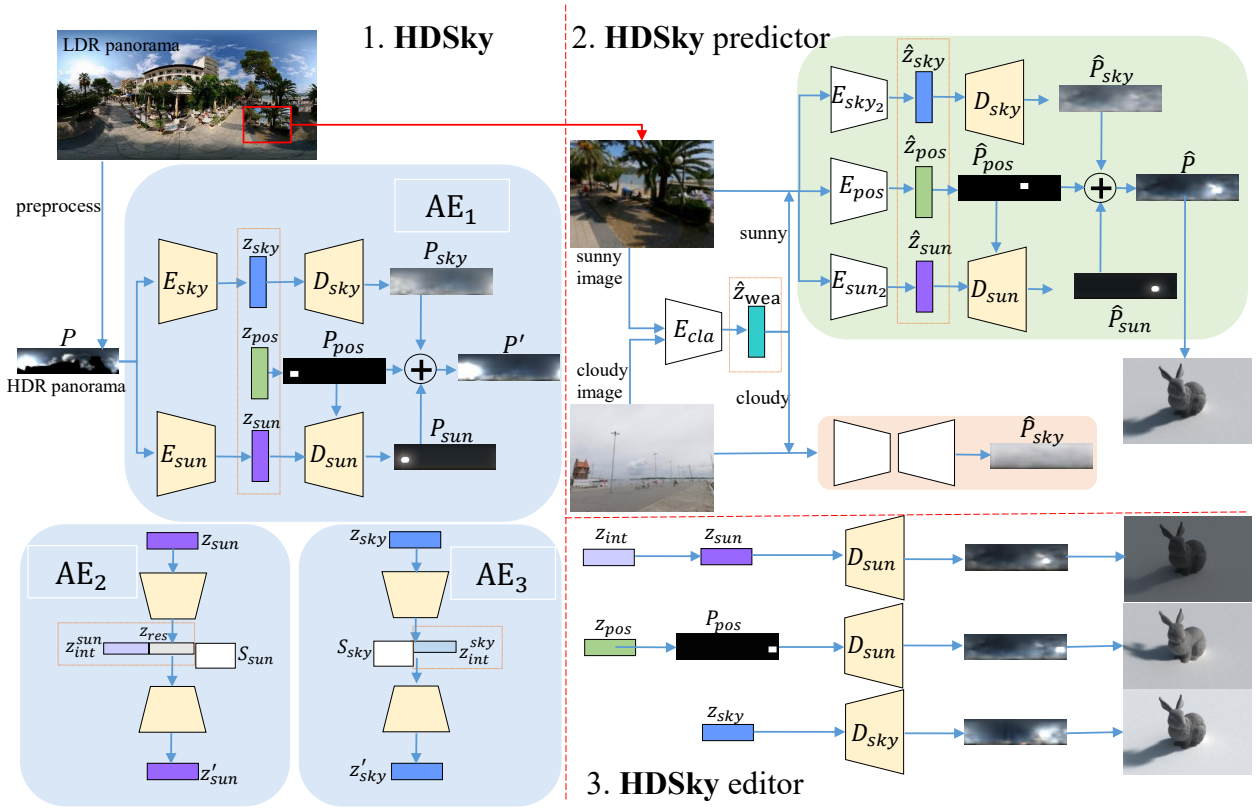

Figure 2. The proposed architecture of our method. We train two well-designed networks (AE_1 and AE_2) to compress and disentangle an input outdoor panorama into several independent and meaningful factors, *i.e.*, HDSky. This disentangled low-dimensional representation facilitates the prediction of all-weather sky information from a single FoV-limited image, improving the accuracy of prediction and enabling consistent shading in AR applications. It also makes outdoor illumination editing more convenient than previous entangled representations.

To enable more intuitive editing, we further disentangle the sun intensity from the sun vector using the second autoencoder AE_2 and disentangle the sky intensity from the sky vector with the third autoencoder AE_3. As shown in the bottom-left corner of Fig. 2, with the sun vector z_{sun} as input, AE_2 can generate the sun intensity vector z_{int}^{sun}, the sun shape S_{sun} and the residual vector z_{res} which captures the remaining variability of z_{sun}. AE_3 can disentangle the sky intensity z_{int}^{sky} and the sky shape S_{sky} from z_{sky}. Once AE_2 and AE_3 are trained, the sun intensity vector z_{int}^{sun} and the sky intensity vector z_{int}^{sky} can be edited directly to modify the sun intensity and the sky intensity of the original (or predicted) panorama. The details of the networks are presented in the supplementary material.

For cloudy panoramas, we train another two autoencoders to perform compression and the disentanglement of the sky intensity, respectively. The supplementary material presents more details.

4.2. Loss functions

To achieve hierarchical disentanglement, AE_1, AE_2 and AE_3 are trained sequentially. We train AE_1 with the objective function:

$$\mathcal{L}_{AE_1} = \alpha \mathcal{L}_{recon} + \mathcal{L}_{sun} + \mathcal{L}_{info}, \tag{2}$$

where we set α to 10 empirically. The reconstruction loss $\mathcal{L}_{recon}$ is used to measure the similarity between the reconstructed panorama and the original one with the L_1 norm. To enforce the sun panorama $\mathcal{P}_{sun}$ only learn the sun information, we design the sun loss $\mathcal{L}_{sun}$ to make the value of the sky area in $\mathcal{P}_{sun}$ tend to 0 with the L_2 norm:

$$\mathcal{L}_{sun} = \|\mathcal{P}_{sun} - \mathcal{P}_{sun} \odot \mathcal{P}_{pos}\|_2. \tag{3}$$

To improve the disentanglement of AE_1, we enforce high mutual information between the vector and the panorama with the information theory. As shown in Fig. 3, two networks E_{sky}^i and E_{sun}^i are utilized to induce z_{sky} and z_{sun} to capture the sky and sun information, respectively:

$$\begin{aligned}\mathcal{L}_{info} = &\max \mathbb{E}_{z_{sky}} \left[\log E_{sky}^i \left(z_{sky}|\mathcal{P}_{sky}\right)\right] + \\ &\max \mathbb{E}_{z_{sun},z_{pos}} \left[\log E_{sun}^i \left(z_{sun}|\mathcal{P}_{sun}\right)\right].\end{aligned} \tag{4}$$

Here, the notation $E(z|P)$ means that the encoder E accepts a panorama $\mathcal{P}$ and outputs a latent code z. E_{sky}^i is

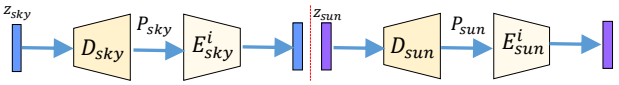

Figure 3. Information theory is used in disentangling outdoor illumination.

used to generate z_{sky} from $\mathcal{P}_{sky}$. The L_2 norm is used to minimize the difference between z_{sky} generated by E_{sky} and the reconstruction generated by E^i_{sky}.

Once AE_1 is trained, we can train AE_2 and AE_3, which is presented in the supplementary material.

4.3. Training details

To train AE_1, 8,466 HDR panoramas from the SUN360-HDR dataset [8] and 4,484 HDR panoramas from the Laval sky dataset [15] are used. Specifically, 9,300 sunny HDR panoramas together with the corresponding sun positions are used to train the complete AE_1. The input HDR panoramas are stored in the latitude-longitude format with a reduced resolution of 32×128 in RGB of the up hemisphere and compressed into log space. In addition, we use 9,300 sun vectors and 9,300 sky vectors learned by AE_1 to train AE_2 and AE_3 respectively.

Once HDSky is trained, we run it on all panoramas in the SUN360-HDR dataset [8], which is obtained by converting LDR panoramas of the SUN360 dataset [32] to HDR, to generate the corresponding disentangled vectors. These vectors together with the images extracted from the SUN360 dataset [32] are then used as training examples for further applications.

5. Estimating illumination from a single image

With HDSky, we are able to achieve higher accuracy in predicting outdoor illumination from a single image than the previous method based on conflated representation [8]. In this section, we describe the overall framework of HDSky predictor to hierarchically estimate all-weather sky information from an FoV-limited image in detail.

5.1. Network architecture

In our hierarchical framework, a classification network E_{cla} is first used to classify the input image as either sunny or cloudy (z_{wea}). The details of the classification network are presented in the supplementary material. We then can separately predict the HDSky illumination of sunny images and cloudy images.

For sunny images, two networks are used to predict the sky vector $\hat{z}_{sky}$ and the sun vector $\hat{z}_{sun}$, as shown in the top-right corner of Fig. 2. The two vectors are then transformed into the sky panorama $\hat{\mathcal{P}}_{sky}$ and sun panorama $\hat{\mathcal{P}}_{sun}$ with the trained decoders D_{sky} and D_{sun} of HDSky. During the transformation, the estimated sun position (discussed later) is used to recover $\hat{\mathcal{P}}_{sun}$. Subsequently, we can obtain the full panorama $\hat{\mathcal{P}}$ by fusing $\hat{\mathcal{P}}_{sky}$ and $\hat{\mathcal{P}}_{sun}$ with Eq. 1.

As aforementioned, sun position is used to recover the sun panorama. We therefore estimate the sun position of sunny images using the network E_{pos} with a pre-trained DenseNet-161 [11] architecture. The difference is that the last layer of our E_{pos} is a fully connected layer of 160 nodes. By discretizing the sky hemisphere into 160 bins (5 for elevation and 32 for azimuth), E_{pos} outputs a probability map over possible sun positions [9]. We then select the position with the maximum probability as the sun. The KL divergence loss is used to train the network. See more details in the supplementary material.

Compared with sunny images, we leverage a single network to predict the sky vector $\hat{z}_{sky}$ of a cloudy image. Once $\hat{z}_{sky}$ is obtained, we can directly transform it into the sky panorama $\hat{\mathcal{P}}_{sky}$ (see the orange block in Fig. 2).

5.2. Loss functions

To train E_{sky_2} and E_{sun_2} for sunny images, we design the vector loss $\mathcal{L}_{z\_sunny}$ to measure the similarity between the predicted vectors with the original ones using the L_1 norm. Furthermore, we add the L_1 loss ($\mathcal{L}_{P\_sunny}$) on the recovered panoramas to capture the sky color and strong sun intensities. In all, the objective function for training the two networks is:

$$\mathcal{L}_{sunny} = \delta \mathcal{L}_{z\_sunny} + \mathcal{L}_{P\_sunny}, \tag{5}$$

where δ is set to 1×10^6 to improve the accuracy of the predicted vectors. For cloudy images, we also leverage the L_1 loss on the sky vector and sky panorama to optimize the cloudy network, where the weight distribution is the same as the above objective function on sunny images.

5.3. Training details

To train the networks in this stage, we prepare a large number of FoV-limited images. Inspired by [9], we extract 7 FoV-limited images with a resolution of 320×240 from each LDR panorama of the SUN360 dataset [32]. 42,056 sunny images are used to train E_{sky_2} and E_{sun_2} for 11 epochs. Besides, the network E_{pos} for estimating the sun position is trained on the same 42,056 sunny images. Convergence of E_{pos} is obtained after 16 epochs. 15,750 cloudy images are employed to train the network which is used to predict the cloudy illumination for 10 epochs. These networks are all trained using the Adam optimizer [14] with $\beta = (0.5, 0.999)$ and the same learning rate of 0.0001.

6. Outdoor illumination editing

Our HDSky achieves the desired disentanglement and can be used in two applicable occasions. In the compression stage, the disentangled vectors obtained by HDSky can be directly edited to generate more panoramas that conform to the physical laws, thereby expanding the outdoor HDR

	Sunny/50		Cloudy/50	
	Panoramas	Renders	Panoramas	Renders
SkyNet [8]	1.532	0.072	0.058	0.030
HDSky	**0.569**	**0.025**	**0.050**	**0.017**
-$\mathcal{L}_{info}$	0.785	0.028	–	–
-$\mathcal{L}_{sun}$	0.757	0.033	–	–
-$\mathcal{P}_{pos}$	0.774	0.026	–	–

Table 1. (Top) Quantitative comparison of panorama reconstruction between HDSky and SkyNet [8]. (Bottom) Ablation studies. RMSE is used to measure the reconstruction quality (↓ better).

Figure 4. Visual comparison of panorama reconstruction between our HDSky and SkyNet [8]. The errors on the bottom left show the higher accuracy of our reconstruction.

panoramas. In the lighting prediction stage, if the predicted lighting of our HDSky predictor deviates from the ground-truth lighting in AR rendering, users can employ HDSky editor to conveniently adjust the predicted lighting to ensure the consistency of the rendering.

Our HDSky editor provides explicit parameters for users to interact with. Specifically, the sun intensity vector generated by AE_2 of HDSky can be modified to intuitively edit the sun intensity of the panorama predicted by HDSky predictor. By directly modifying the sun position during the recovery of the panorama (see the top-right corner of Fig. 2), the position of the sun in the predicted panorama can be smoothly changed. In addition, the sky intensity of the predicted panorama can also be edited with explicit parameters.

7. Experiments

In this section, we first compare our HDSky against SkyNet [8] which is the state-of-the-art data-driven sky model in reconstructing the outdoor panoramas. Then, the performance of HDSky predictor in outdoor illumination prediction is evaluated compared with SkyNet [8] and the classical analytical model: the method of Hold-Geoffroy *et al*. [9]. Finally, we evaluate HDSky editor and the disentanglement performance of HDSky. Ablation studies are also conducted.

	Sunny/100		Cloudy/100	
	Panoramas	Renders	Panoramas	Renders
[9]	20.52	0.633	21.86	1.155
SkyNet [8]	10.69	0.444	0.314	0.101
Ours	**9.756**	**0.172**	**0.247**	**0.030**

Table 2. Quantitative comparison in terms of RMSE of different methods in outdoor illumination prediction. Our HDSky predictor performs significantly better than other methods. The accuracy of the method of Hold-Geoffroy *et al*. [9] severely degrades in the cloudy conditions due to the limitation of the analytical model.

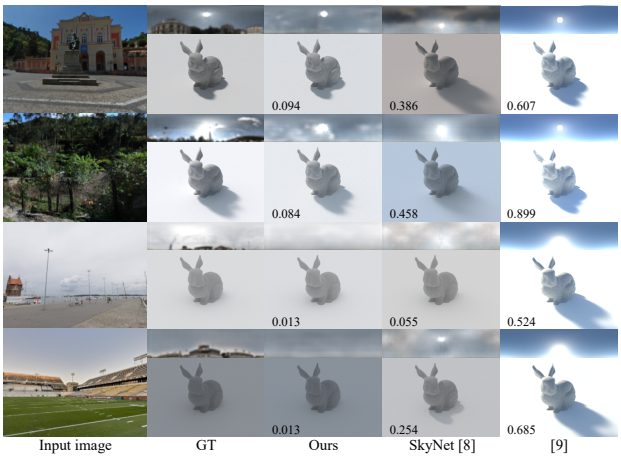

Figure 5. Visual comparison of outdoor illumination prediction between different methods. The RMSE errors are reported in the bottom left of rendered images.

7.1. Reconstruction quality of HDSky

We quantitatively evaluate the reconstruction quality of HDSky and compare it with SkyNet [8] which utilizes a single latent vector to compress the outdoor illumination. Two test sets from the Laval sky dataset [15] and the SUN360-HDR dataset [8] are employed. One contains 50 sunny HDR panoramas (Sunny/50), and the other contains 50 cloudy HDR panoramas (Cloudy/50). The rendered images synthesized with these panoramas are also used. We then adopt the RMSE on these panoramas and rendered images to quantify the performance of different methods. As shown in Table 1, our HDSky performs significantly better than SkyNet [8] due to the disentangled representation of outdoor illumination.

Fig. 4 shows some visual examples from the SUN360-HDR dataset [8]. The errors in terms of RMSE are shown in the bottom left of the rendered images. Overall, our HDSky achieves higher reconstruction quality than SkyNet [8] with more accurate sun intensity, shading and shadows. Since HDSky separates the sky and sun information, the mutual influence between the two factors is eliminated. For example, HDSky can more accurately reconstruct the panorama of the rightmost column where the sun is blocked by buildings and the sky dominates the full panorama.

Figure 6. Visual comparison of different methods for different viewpoints from the same panorama of the SUN360 dataset [32]. Compared with other methods [8, 9], our HDSky predictor generates smoother transitions from one viewpoint to another.

7.2. Evaluation of HDSky predictor

To evaluate the performance of HDSky predictor to predict the disentangled illumination from a single image, we utilize images extracted from the SUN360 dataset [32]. We construct another two test sets: 100 sunny images (Sunny/100) and 100 cloudy images (Cloudy/100). The quantitative comparison of different methods for predicting outdoor lighting from a single image is listed in Table 2. The results reveal that our HDSky predictor outperforms its competitors [9, 8] on the predicted panoramas and the rendered images. Because the analytical models fail to fully capture the complexities of real-world lighting conditions with few parameters, the RMSE of the method of Hold-Geoffroy *et al*. [9] is much higher than our HDSky predictor, especially on the cloudy images. Our HDSky predictor achieves better performance than SkyNet [8] due to the disentangled representation.

Fig. 5 shows the qualitative comparison between different methods under different weather conditions. Our HDSky predictor estimates outdoor illumination with higher accuracy compared with its competitors [9, 8]. The method of Hold-Geoffroy *et al*. [9] generates very intense sun for both sunny images and cloudy images. Because SkyNet [8] uses a single vector to represent the outdoor illumination, the sun and sky are entangled, which reduces the accuracy of outdoor illumination prediction. For example, SkyNet [8] predicts inaccurate sun intensity and sun elevation in the first two rows of Fig. 5. Moreover, an improper sun is predicted by SkyNet [8] from a cloudy image in the last row. In contrast, our HDSky predictor produces plausible shading and shadows under different weather conditions. Furthermore, our HDSky predictor generates accurate outdoor lighting and provides consistent shading and shadows as the viewpoint changes (see Fig. 6). Also, Fig. 7 shows that the estimated lighting intensity of our HDSky predictor can provide plausible shadows on real-world images under different weather conditions.

Figure 7. Demonstration of virtual object insertion. Benefited from our HDSky, different virtual objects (the horse and dragon) are inserted into real scenes with consistent shading and shadows.

7.3. Evaluation of HDSky Editor

We further evaluate how well HDSky disentangles each factor (the sky vector z_{sky}, the sun position z_{pos} and the sun vector z_{sun}) and generates realistic panoramas. Our disentanglement of each factor on HDR panoramas is shown in Fig. 8. For each subfigure, the panoramas in the top row and leftmost column (with red boxes) are reconstructed panoramas. The specific factors taken from each reconstructed panorama are indicated in the top-left corner. For example, in (a), the sky is taken from the top row, while the sun position and the sun are taken from the leftmost column. We can change the sky, the sun position and the sun of a reconstructed panorama by varying (a) z_{sky}, (b) z_{pos} and (c) z_{sun} to obtain the synthetic panoramas in the 3×3 area. It is impressive that our HDSky can generate realistic and diverse panoramas by modifying the latent vectors.

Our HDSky editor provides explicit parameters for users to edit the predicted panorama. The visually smooth transitions are shown in Fig. 9, where the predicted lighting of the image in the top-left corner is marked with red boxes. We can directly modify the sun intensity and smoothly change the sun intensity of the predicted panorama without affecting the sky information (the first row). In addition, changing the azimuth and elevation of the sun (the last two rows) will not affect any other information due to the complete disentanglement of the sun position from the outdoor panorama.

7.4. Ablation studies

AE_1 of HDSky utilizes three key components: the loss of the information theory $\mathcal{L}_{info}$, the sun loss $\mathcal{L}_{sun}$ and the sun position map $\mathcal{P}_{pos}$. To show their effectiveness, we conduct some ablation studies by removing each of them from our complete AE_1 . The bottom part in Table 1 proves that all components are necessary for training AE_1. Otherwise, disentanglement cannot be learned properly and the reconstruction quality will reduce. Several visual examples are shown in Fig. 10. The results show that the rendered images of our complete AE_1 display more accurate shadows under different solar intensity.

For a fair comparison with SkyNet [8], we set the sum of

(a) Varying z_{sky} (sky) (b) Varying z_{pos} (sun position) (c) Varying z_{sun} (sun)

Figure 8. Varying a single lighting factor. Reconstructed panoramas are indicated with red boxes. The reconstructed panoramas on the left/top provide two/one factors for each synthetic panorama. The center 3×3 panoramas are synthetic panoramas with changed factors. The panoramas in the 2 subfigures on the left are from the Laval sky dataset [15], and the rightmost are from the SUN360-HDR dataset [8].

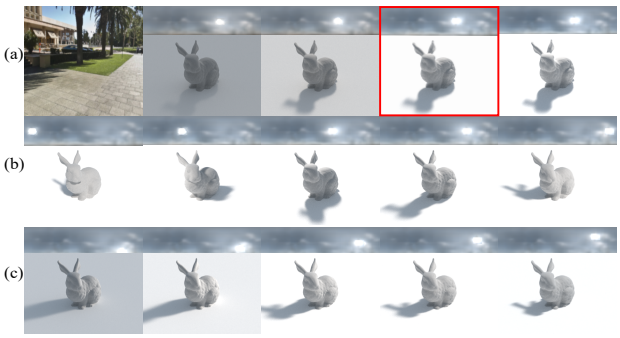

Figure 9. Examples of intuitive edits of our HDSky editor. (a) The rendered image in the red boxes is generated with the predicted panorama of the given image. Changing the sun intensity of the predicted panorama generates smooth transitions. We can also smoothly edit the sun azimuth (b) and sun elevation (c) without affecting any other information.

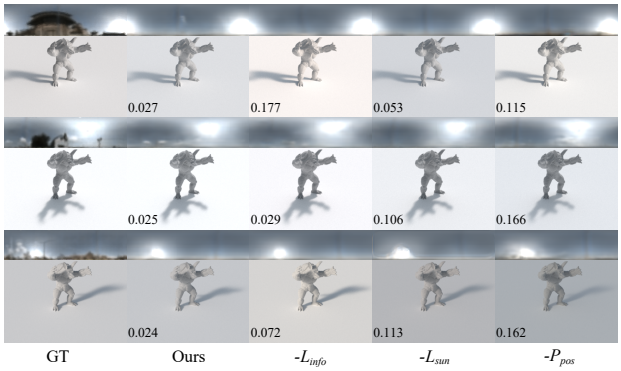

Figure 10. Visual comparison of ablation studies for the key components of AE_1.

the dimensions of all disentangled vectors compressed by AE_1 to 64. Among them, the dimensions of the weather vector and the sun position are 1 and 2, respectively. In our current implementation, the dimensions of the sky vector z_{sky} and the sun vector z_{sun} are set to 8 and 53, respectively. To prove the effectiveness of this allocation, we train another 4 pipelines with different dimensional ratios of z_{sky} and z_{sun} (16:45, 24:37, 32:29, 48:13). The RMSE metric is adopted to evaluate the accuracy of these pipelines in predicting outdoor lighting from a single image. The errors on 300 randomly selected sunny images reported in Table 3 show that the ratio 8:53 leads to the best estimation quality.

z_{sky}: z_{sun}	8:53	16:45	24:37	32:29	48:13
RMSE	**10.28**	11.84	12.31	14.18	10.51

Table 3. Ablation studies of different dimensional ratios of z_{sky} and z_{sun} of our approach.

7.5. Limitation and future work

Despite its success, our HDSky suffers from the following limitation. As shown in Fig. 5, while our HDSky predictor estimates more accurate lighting than other methods, it sometimes cannot accurately predict the sun shape (the third row). In the future, we will explore the relationship between the shape of the sun and other factors (*e.g.*, the sun position and sun intensity) to solve the problem. It is also interesting to develop a unified representation to handle both outdoor and indoor illumination.

8. Conclusion

In this paper, we have proposed a hierarchical disentangled sky model for outdoor illumination to compress each sunny panorama into several meaningful vectors. A hierarchical framework is designed to estimate the individual latent vectors from every sunny image and generate the outdoor illumination based on the sky model. For cloudy images, we only predict the sky information. We further show how this model generates realistic and diverse outdoor panoramas and provides explicit parameters to intuitively edit the predicted outdoor illumination. The sky model allows us to represent outdoor illumination more faithfully and supports all-weather conditions. As demonstrated via extensive quantitative and qualitative evaluations, our hierarchical disentangled sky model outperforms previous analytical models and data-driven models.

9. Acknowledgement

We would like to thank the anonymous reviewers for their valuable comments. This work was supported by NSFC (62032011, 61972194 and 61772257).

References

[1] Guha Balakrishnan, Amy Zhao, Adrian V. Dalca, Fredo Durand, and John Guttag. Synthesizing images of humans in unseen poses. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8340–8348, 2018.

[2] Jonathan T. Barron and Jitendra Malik. Intrinsic scene properties from a single RGB-D image. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 17–24, 2013.

[3] Jonathan T. Barron and Jitendra Malik. Shape, illumination, and reflectance from shading. *IEEE Trans. Pattern Anal. Mach. Intell.*, 37(8):1670–1687, 2015.

[4] Dan A. Calian, Jean-François Lalonde, Paulo F. U. Gotardo, Tomas Simon, Iain A. Matthews, and Kenny Mitchell. From faces to outdoor light probes. *Comput. Graph. Forum*, 37(2):51–61, 2018.

[5] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016*, pages 2172–2180, 2016.

[6] Dachuan Cheng, Jian Shi, Yanyun Chen, Xiaoming Deng, and Xiaopeng Zhang. Learning scene illumination by pairwise photos from rear and front mobile cameras. *Comput. Graph. Forum*, 37(7):213–221, 2018.

[7] Stamatios Georgoulis, Konstantinos Rematas, Tobias Ritschel, Efstratios Gavves, Mario Fritz, Luc Van Gool, and Tinne Tuytelaars. Reflectance and natural illumination from single-material specular objects using deep learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(8):1932–1947, 2018.

[8] Yannick Hold-Geoffroy, Akshaya Athawale, and Jean-François Lalonde. Deep sky modeling for single image outdoor lighting estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6920–6928, 2019.

[9] Yannick Hold-Geoffroy, Kalyan Sunkavalli, Sunil Hadap, Emiliano Gambaretto, and Jean-François Lalonde. Deep outdoor illumination estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2373–2382, 2017.

[10] Lukas Hosek and Alexander Wilkie. An analytic model for full spectral sky-dome radiance. *ACM Trans. Graph.*, 31(4):95:1–95:9, 2012.

[11] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2261–2269, 2017.

[12] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4396–4405, 2019.

[13] Karsch Kevin, Kalyan Sunkavalli, Sunil Hadap, Nathan Carr, Hailin Jin, Rafael Fonte, Michael Sittig, and David A. Forsyth. Automatic scene inference for 3d object compositing. *ACM Trans. Graph.*, 33(3):32:1–32:15, 2014.

[14] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations*, 2015.

[15] Jean-François Lalonde, Louis-Philippe Asselin, Julien Becirovski, Yannick Hold-Geoffroy, Mathieu Garon, Marc-André Gardner, and Jinsong Zhang. The laval hdr sky database, 2015.

[16] Jean-François Lalonde, Alexei A. Efros, and Srinivasa G. Narasimhan. Estimating the natural illumination conditions from a single outdoor image. *Int. J. Comput. Vision*, 98(2):123–145, June 2012.

[17] Jean-François Lalonde and Iain Matthews. Lighting estimation in outdoor image collections. In *International Conference on 3D Vision*, pages 131–138, 2014.

[18] Chloe LeGendre, Wanchun Ma, Graham Fyffe, John Flynn, Laurent Charbonnel, Jay Busch, and Paul Debevec. Deeplight: Learning illumination for unconstrained mobile mixed reality. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5911–5921, 2019.

[19] Chloe LeGendre, Wanchun Ma, Rohit Pandey, Sean Ryan Fanello, Christoph Rhemann, Jason Dourgarian, Jay Busch, and Paul E. Debevec. Learning illumination from diverse portraits. *CoRR*, abs/2008.02396, 2020.

[20] Yuheng Li, Krishna K. Singh, Utkarsh Ojha, and Yong J. Lee. Mixnmatch: Multifactor disentanglement and encoding for conditional image generation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8036–8045, 2020.

[21] Daquan Liu, Chengjiang Long, Hongpan Zhang, Hanning Yu, Xinzhi Dong, and Chunxia Xiao. Arshadowgan: Shadow generative adversarial network for augmented reality in single light scenes. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8136–8145, 2020.

[22] Stephen Lombardi and Ko Nishino. Reflectance and illumination recovery in the wild. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(1):129–141, 2016.

[23] Lukáš Hošek and Alexander Wilkie. Adding a solar-radiance function to the hošek-wilkie skylight model. *IEEE Computer Graphics and Applications*, 33(3):44–52, 2013.

[24] Xi Peng, Xiang Yu, Kihyuk Sohn, Dimitris N. Metaxas, and Manmohan Chandraker. Reconstruction-based disentanglement for pose-invariant face recognition. In *IEEE International Conference on Computer Vision*, pages 1632–1641, 2017.

[25] Richard Perez, Robert Seals, and Joseph Michalsky. All-weather model for sky luminance distribution—preliminary configuration and validation. *Solar Energy*, 50(3):235 – 245, 1993.

[26] Arcot J. Preetham, Peter Shirley, and Brian Smits. A practical analytic model for daylight. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, page 91–100, 1999.

[27] Konstantinos Rematas, Tobias Ritschel, Mario Fritz, Efstratios Gavves, and Tinne Tuytelaars. Deep reflectance maps. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4508–4516, 2016.

[28] Pinar Satilmis, Thomas Bashford-Rogers, Alan Chalmers, and Kurt Debattista. A machine-learning-driven sky model.

IEEE Computer Graphics and Applications, 37(1):80–91, 2017.

[29] Davoud Shahlaei and Volker Blanz. Realistic inverse lighting from a single 2d image of a face, taken under unknown and complex lighting. In *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, pages 1–8, 2015.

[30] Hyunjung Shim. Faces as light probes for relighting. *Optical Engineering*, 51(7):1 – 8, 2012.

[31] Krishna K. Singh, Utkarsh Ojha, and Yong J. Lee. Finegan: Unsupervised hierarchical disentanglement for fine-grained object generation and discovery. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6483–6492, 2019.

[32] Jianxiong Xiao, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. Recognizing scene viewpoint using panoramic place representation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2695–2702, 2012.

[33] Jinsong Zhang, Kalyan Sunkavalli, Yannick Hold-Geoffroy, Sunil Hadap, Jonathan Eisenmann, and Jean-François Lalonde. All-weather deep outdoor lighting estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 10158–10166, 2019.

[34] Hao Zhou, Jin Sun, Yaser Yacoob, and David W. Jacobs. Label denoising adversarial network (ldan) for inverse lighting of faces. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6238–6247, 2018.