AN OPTICAL PHYSICS INSPIRED CNN APPROACH FOR INTRINSIC IMAGE DECOMPOSITION

Harshana Weligampola[†] *Gihan Jayatilaka*[⋆] *Suren Sritharan*[†]
Parakrama Ekanayake[⋆] Roshan Ragel[⋆] Vijitha Herath[⋆] Roshan Godaliyadda[⋆]

[⋆] Faculty of Engineering, University of Peradeniya, Peradeniya [20400], Sri Lanka
[†] Faculty of IT and Computing, Sri Lanka Technological Campus, Padukka [10500], Sri Lanka

ABSTRACT

Intrinsic Image Decomposition is an open problem of generating the constituents of an image. Generating reflectance and shading from a single image is a challenging task specifically when there is no ground truth. There is a lack of unsupervised learning approaches for decomposing an image into reflectance and shading using a single image. We propose a neural network architecture capable of this decomposition using physics-based parameters derived from the image. Through experimental results, we show that (a) the proposed methodology outperforms the existing deep learning-based IID techniques and (b) the derived parameters improve the efficacy significantly. We conclude with a closer analysis of the results (numerical and example images) showing several avenues for improvement.

Index Terms— intrinsic image decomposition (IID), convolutional neural networks (CNN), Phong illumination model.

1. INTRODUCTION

Intrinsic Image Decomposition (IID) is the problem of reverting an image into its building blocks (reflectance, shading, surface normal, etc.). The reflectance depends on the material properties such as shape and color, while the shading contains information about the lighting of the environment. Information such as the geometry of the objects, shadows, directed illumination, and other ambient lights can be derived from the shading. Thus, extracting intrinsic features of an image is essential for various computer vision tasks. For example, using the reflectance (albedo) of an image, segmentation can be done more accurately invariant of the lighting condition [1]. Further, tasks such as image relighting, gamma correction, and recoloring can be easily accomplished using the reflectance and shading information. Therefore, it is essential to identify these intrinsic properties to guarantee the robustness of computer vision algorithms.

Many attempts have been made to decompose images into meaningful constituents. Most notably, The Retinex Theory [2] is a biologically-motivated theory based on the color constancy property of the human visual system (HVS), and it pioneered the field. Image decomposition based on this model attempts to generate reflectance and illumination maps. This has been proven useful in applications such as lighting enhancement [3, 4]. The major drawback of retinex models is their agnosticity to object surface geometries and the complicated physical phenomena related to light reflection.

The Phong illumination model [5] proposed a better theory on imaging based on optical physics. A large body of work has been built upon this including better 3D rendering [6] using shading techniques. This model considers the light reflected by an object as a combination of 3 types – ambient, specular, and diffused. Our work is based on this model. We propose a novel Reflectance approximation map to train the neural network and a physics-based loss function to learn intrinsic properties in an image. We combine these losses and extracted feature maps to train a neural network in an unsupervised manner. Through experimentation, we show that our model is more robust at decomposing images under a diverse set of scenes and lighting conditions compared to the existing deep learning approaches based on numerical metrics as well as sample images.

2. RELATED WORK

IID has been attempted as a sequential algorithm, optimization problem, and a trainable neural networks problem. Most of these approaches try to decompose in a way that reconstruction from the decomposed components is consistent with the original image.

Classical models employ well defined mathematical models to formulate the problem as an optimization problem, and thereby decompose the image [7–11]. The main drawback of

©2021 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

This paper is under review for ICIP 2021.

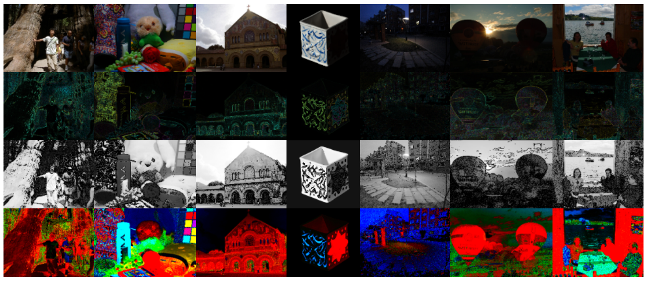

Fig. 1: Comparison of different images (first row) with computed RRG, SG, and RAM (following rows in order)

this class of models is the limitations of the mathematical definitions to capture the wide variety of imaging conditions and image capture artifacts.

Deep learning approaches try to build a model that incorporates the desirable properties of the image through loss functions. This definition is properly adapted to a wider variety of scenes by the use of large datasets [12–15]. The limitations of these models include overfitting to data, blackboxness, and actual decomposition diverging away from the physics.

3. METHODOLOGY

3.1. Image model

We use the Phong Reflectance model [5], which is widely used to formulate image modeling. It describes a point in an image as a combination of ambient, diffuse, and specular highlights. For each light source in the scene, $i_d(\lambda)$ and $i_s(\lambda)$ are defined as the diffuse and specular intensity distribution components of the light source, where λ is the wavelength of light. For multiple light sources ($\mathbf{L}$) the intensity of light reflected from a point p that is represented in the image can be defined as,

$$I_p = \int_\lambda k_a r_p(\lambda) i_a(\lambda) + \sum_{\hat{L}^{(n)} \in \mathbf{L}} \{k_d r_p(\lambda)[\hat{L}^{(n)}.\hat{N}_p] i_d^{(n)}(\lambda) \\ + k_s s_p(\lambda)[\hat{R}^{(n)}.\hat{V}]^\gamma i_s^{(n)}(\lambda)\} d\lambda \tag{1}$$

where k_a, k_d, and k_s are the ambient, diffuse, and specular coefficients. $r_p(\lambda)$ is the diffuse spectral reflectance and $s_p(\lambda)$ is the specular spectral reflectance at point p. $\hat{L}_n$ is the direction vector from a point on the surface to the light direction. $\hat{N}_p$ is the normal at point p. $\hat{R}^{(n)}$ is the direction in which a perfectly reflected ray of light would travel. $\hat{V}$ is the direction pointing to the viewer. Note that a hat ($\hat{}$) represents that the parameter is a vector.

We assume that the specular term is negligible in most points on the surface. Then, considering a narrow band (λ_c) we can reduce Eq. (1) to,

$$I_p(\lambda_c) = r_p(\lambda_c)[k_a i_a(\lambda_c) + \sum_{\hat{L}^{(n)} \in \mathbf{L}} k_d[\hat{L}^{(n)}.\hat{N}_p] i_d^{(n)}(\lambda_c)] \tag{2}$$

Assuming that only one light source exists and that the ambient illumination is constant, we can write Eq. (2) after eliminating constant k_d as,

$$I_p(\lambda_c) = r_p(\lambda_c)[\hat{L}.\hat{N}_p] i_d(\lambda_c) \tag{3}$$

In Eq. (3), we can define reflectance as $\mathbf{R} = [r_p(\lambda_c)]$ and shading as $\mathbf{S} = [[\hat{L}.\hat{N}_p] i_d(\lambda_c)]$ where $\mathbf{R}$ and $\mathbf{S}$ are matrices where each element corresponds to a pixel p. Theoretically, object shape feature (normal) is included in the shading $\mathbf{S}$. But practically some of the shape features propagate to the reflectance due to the associative property of element-wise multiplication of shading and reflectance.

3.2. Reflectance ratio gradient (RRG)

Consider two narrow-band channels λ_a and λ_b. Substituting them in Eq. (3) we get two images in different wavelengths where each pixel is given by $I_p(\lambda_a)$ and $I_p(\lambda_b)$. As in [15], we consider the natural logarithm of the ratio between $I_p(\lambda_a)$ and $I_p(\lambda_b)$.

$$\mathcal{J}_p(\lambda_a, \lambda_b) = log\left(\frac{I_p(\lambda_a)}{I_p(\lambda_b)}\right) = log\left(\frac{r_p(\lambda_a) i_d(\lambda_a)}{r_p(\lambda_b) i_d(\lambda_b)}\right) \tag{4}$$

For two neighboring pixels p_1 and p_2, we can assume that light intensities for a given wavelength are the same for both of these pixels (ignoring shadows). Thus, the gradient of Eq. (4) can be written as,

$$\nabla\mathcal{J}(\lambda_a, \lambda_b) = \nabla log\left(\frac{r(\lambda_a)}{r(\lambda_b)}\right) \tag{5}$$

From Eq. (5) we can show that for two given narrow wavelength bands of light, we can find the gradient of object reflectance ratio of corresponding two wavelengths. Images have red (R), green (G), and blue (B) channels corresponding to the three wavelength bands λ_R, λ_G, and λ_B. Using Eq. (5) we get, $\nabla\mathcal{J}(\lambda_R, \lambda_G) = \nabla log\left(\frac{r(\lambda_R)}{r(\lambda_G)}\right)$, $\nabla\mathcal{J}(\lambda_R, \lambda_B) = \nabla log\left(\frac{r(\lambda_R)}{r(\lambda_B)}\right)$, $\nabla\mathcal{J}(\lambda_B, \lambda_G) = \nabla log\left(\frac{r(\lambda_B)}{r(\lambda_G)}\right)$. These gradients are calculated for a local neighborhood using the derivative of 2D Gaussian filter. These three gradients are referred to as the Reflectance Ratio Gradient (RRG). We can use RRG to identify the boundaries of the uniform reflectance in an image. Let f_{RRG} be the function that converts an image to RRG.

3.3. Reflectance Approximation Map (RAM)

When $r(\lambda_c)$ for all c wavelengths in Eq. (3) are almost equal (surfaces with a shade of white), RRG will be close to zero.

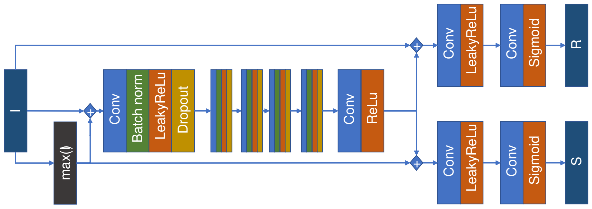

Fig. 2: Illustration of the neural network architecture.

Therefore, we define a reflectance approximation map that gives the likelihood of a particular channel being significant for reflectance. To define this map we first clip values from 0 to 1 in Eq. (4) which is given by $\overline{\mathcal{J}}_p(\lambda_a, \lambda_b)$. Then RAM can be given as follows,

$$M_{RAM} = \left[m_p^{(c)}\right] = \begin{cases} (\overline{\mathcal{J}}_p(\lambda_R, \lambda_G) + \overline{\mathcal{J}}_p(\lambda_R, \lambda_B))/2 & \text{if } c = R \\ (\overline{\mathcal{J}}_p(\lambda_G, \lambda_R) + \overline{\mathcal{J}}_p(\lambda_G, \lambda_B))/2 & \text{if } c = G \\ (\overline{\mathcal{J}}_p(\lambda_B, \lambda_G) + \overline{\mathcal{J}}_p(\lambda_B, \lambda_R))/2 & \text{if } c = B \end{cases} \tag{6}$$

We can use Eq. (6) to identify whether the predicted reflectance is accurate. For example, if $m_p^{(R)}$ is greater than $m_p^{(G)}$ or $m_p^{(B)}$, we can say that the red channel of the albedo should be significant. Note that Eq. (6) does not give actual reflectance values, but the likelihood of the reflectance. Let f_{RAM} be the function that converts an image to RAM.

3.4. Shading Gradient (SG)

Consider a narrow band channel wavelength λ_a. Then take the natural logarithm of $I_p(\lambda_a)$.

$$\mathcal{K}_p(\lambda_a) = log\left(I_p(\lambda_a)\right) \tag{7}$$

If we consider two neighboring pixels p_1 and p_2 with constant reflectance $r_{p_1}(\lambda_a) \approx r_{p_2}(\lambda_a)$, the gradient of Eq. (7) can be written as,

$$\begin{aligned} \nabla\mathcal{K}(\lambda_a) &= \mathcal{K}_{p_1}(\lambda_a) - \mathcal{K}_{p_2}(\lambda_a) \\ &= \nabla log\left([\hat{L}.\hat{N}]\right) \end{aligned} \tag{8}$$

As we can see from Eq. (8), the gradient of log intensity is equal to the gradient of normal.

For each wavelength band, it is valid only for pixels where the reflectance corresponding to the wavelength is equal in the given neighborhood. For example, in the red channel, Eq. (10) is valid only if $r_{p_1}(\lambda_R) \approx r_{p_2}(\lambda_R)$ where p_1 and p_2 are two pixels that are in a given neighborhood. We can approximate that the reflectance for red channel is equal in the given neighborhood if RRG for red channel is low. i.e. $\nabla\mathcal{J}(\lambda_R, \lambda_G)$ and $\nabla\mathcal{J}(\lambda_R, \lambda_B)$ are low. Also for the green channel, if $\nabla\mathcal{J}(\lambda_G, \lambda_B)$ and $\nabla\mathcal{J}(\lambda_G, \lambda_R)$ are very low we can assume that the reflectance for green channel is very low. Similarly, for blue channel, $\nabla\mathcal{J}(\lambda_B, \lambda_G)$ and $\nabla\mathcal{J}(\lambda_B, \lambda_R)$ should be very low. Therefore, we can create a three-channel map using RRG where Eq. (8) is invalid for RGB channels. This map can be given as follows.

$$M_{RRG} = \left[m_p^{(c)}\right] = \begin{cases} (\nabla\mathcal{J}(\lambda_R, \lambda_G) + \nabla\mathcal{J}(\lambda_R, \lambda_B))/2 & \text{if } c = R \\ (\nabla\mathcal{J}(\lambda_G, \lambda_B) + \nabla\mathcal{J}(\lambda_G, \lambda_R))/2 & \text{if } c = G \\ (\nabla\mathcal{J}(\lambda_B, \lambda_G) + \nabla\mathcal{J}(\lambda_B, \lambda_R))/2 & \text{if } c = B \end{cases} \tag{9}$$

For RGB channels, we can get three gradients of normal using wavelength bands λ_R, λ_G and λ_B that are masked by Eq. (9).

$$\nabla\mathcal{K}(\lambda_c) = \begin{cases} \nabla log\left([\hat{L}.\hat{N}]\right) & \text{if } M_{RRG}^{(c)} < 0.1 \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

where the threshold 0.1 is selected arbitrarily. After calculating Eq. (10) for each RGB channel, we can call them as Shading Gradient (SG). Let f_{SG} be the function that converts an image to SG.

3.5. Intrinsic Image decomposition

In this section, we introduce the proposed IID model, its architecture, and loss functions.

The proposed model has one input which is the image and two outputs: a three-channel reflectance map and a single channel shading map. The overall design of the neural network is illustrated in Fig. 2. First, the input image is concatenated with its maximum from RGB channels. This is connected to a sequence of five similar convolution blocks. Each convolution block consist of Reflection padding layer, a 2D Convolution layer with leaky ReLU activation function. The output of the final convolution block is connected to two sets of layers that will output the reflectance and shading. The final output layers are activated using sigmoid activation. The loss function used to train the model has mainly five components. They are given as follows,

$$\mathcal{L} = \alpha_1\mathcal{L}_{recon} + \alpha_2\mathcal{L}_{ss} + \alpha_3\mathcal{L}_{rrg} + \alpha_4\mathcal{L}_{sg} + \alpha_5\mathcal{L}_{ram} \tag{11}$$

where $\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5$ are coefficients that are used to balance the loss function to train the model optimally.

Reconstruction loss is based on the assumption that all reflectance maps are invariable of the lighting condition. Thus, for image $i \in \mathcal{I}$ we should be able to reconstruct the original image from reflectance and shading. Therefore, reconstruction loss is given by, $\mathcal{L}_{recon} = ||\mathbf{R}_i\mathbf{S}_i - \mathbf{I}_i||_1$

Shading smoothness loss ensures that the shading map is smooth where the RRG is smooth. [14] uses the reflectance map generated by the neural network itself. However, this results in color leakage to the shading map, creating a positive feedback loop which leaks texture information to the shading map. Through RRG, we avoid such information leaks from the predicted reflectance map to the shading map, and this loss is given by, $\mathcal{L}_{ss} = ||\nabla\mathbf{S}_i \exp(-10 f_{RRG}(i))||_1$

RRG loss ensures the model learns the representations implied by Eq. (5). Then, RRG loss can be given as follows, $\mathcal{L}_{rrg} = ||f_{RRG}(\mathbf{R}_i) - f_{RRG}(i)||_1$

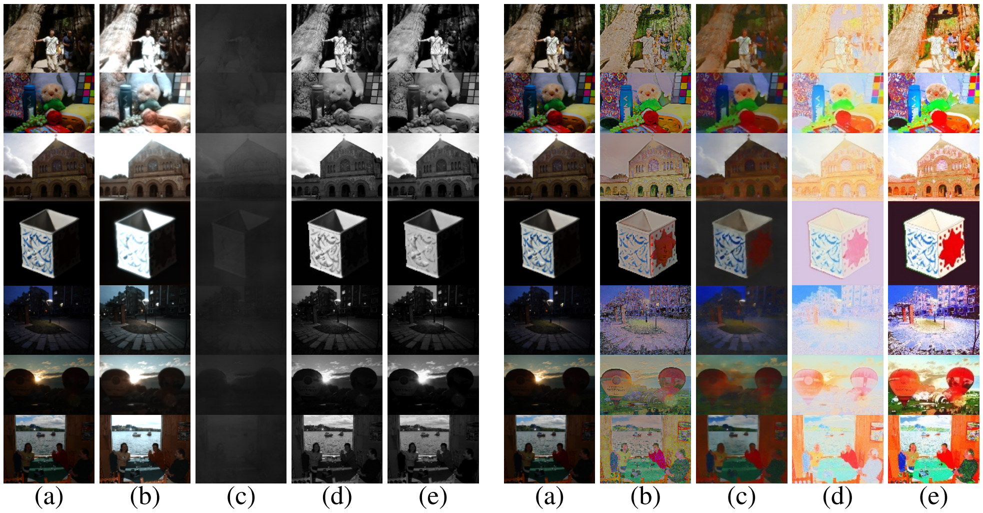

Fig. 3: Comparison of shading (left) and reflectance (right) outputs: (a) Original image, (b)-(d) [12–14], (i) Ours

Table 1: Numerical analysis on reconstructed images, reflectance (R) and shading (S).

Method	LOL dataset				MIT dataset				MIT(R)		MIT(S)	
	RMSE	PSNR	SSIM	NIQE	RMSE	PSNR	SSIM	NIQE	RMSE	PSNR	RMSE	PSNR
Letry et.al	21.87	35.28	**0.96**	7.75	6.67	39.26	**0.99**	12.06	**41.91**	**16.58**	40.88	16.46
CGIntrinsic	63.28	18.95	0.36	**14.78**	40.95	17.36	0.11	**17.47**	48.47	16.28	59.62	12.99
Retinex-net	6.88	34.64	0.90	7.63	3.77	37.85	0.95	14.02	67.39	13.48	37.97	18.54
Ours	**2.00**	**43.12**	0.95	7.63	**1.04**	**41.66**	0.96	14.02	45.90	15.82	**30.54**	**20.14**

SG loss is used to pretain the shape information in the shading map using Eq. (10). First, we reduce the channel dimension of Eq. (10) by element-wise multiplication of each channel. Let's call the channel reduced SG function as f'_{SG}. Then SG loss can be given as, $\mathcal{L}_{sg} = ||(\nabla \log(\mathbf{S_i}) - f'_{SG}(i)) \times f'_{SG}(i)||_1$

RAM loss ensures that the reflectance map is similar to the RAM except for the areas with shades of white. Since these local neighborhoods with shades of white is low in RAM, we define the RAM loss as follows, $\mathcal{L}_{ram} = ||\mathbf{R}_i - f_{RAM}(i)) \times f_{RAM}(i)||_1$

4. EXPERIMENTS

This section presents a series of experiments that were conducted to evaluate the proposed model. We focus mainly on the model's ability to decompose reflectance and shading into relevant categories properly.

The proposed a neural network (described in Section 3.5) contains convolution blocks with 64 filters and 3 × 3 kernels. It is connected to a convolution layer with 32 filters which follows two parallel layers that have 16, 8, 4 filters in each convolution layer as illustrated in Fig. 2. Corresponding values for $\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5$ are $1.0, 0.01, 0.01, 0.0001, 0.1$. The model was implemented using TensorFlow [16]. We trained the neural network with LOL dataset [14] for 100 epochs. 2000 randomly cropped image patches of size 64×64 were fed into the CNN. Each image patch was randomly augmented by flips (horizontal and vertical) and random rotations $(90°, 180°, 270°)$. The model was optimized using adam optimizer [17] set at $\beta = 0.9$ with a learning rate of 0.002 and a decay factor of $e^{-0.01}$ for each epoch.

The sample images in Fig. 3 were randomly selected from multiple datasets to represent a wide variety of sceneries and objects. The decomposed reflectance and shading components using our model and the prior works in [12–14] are shown in Fig. 3. Furthermore, the images reconstructed from these decomposed components (R & S) were evaluated using MSE, NIQE [18], SSIM and, PSNR [19] image quality assessment metrics. These numerical metrics on LOL and MIT datasets are given in Table 1 (columns 1 & 2). In addition, the decomposed components (R, S) were also compared with the ground truth in the case of the MIT dataset (columns 3 & 4) . Through the first 2 columns, we conclude that the proposed model provides better reconstruction in comparison to other deep learning methods. Furthermore, the computed metrics on the decomposed reflectance and shading show that the proposed model consistently performs well.

RMSE : lower is better. PSNR, SSIM, NIQE : higher is better.
Best metric is bold and second best is underlined.

5. CONCLUSIONS

In this paper, we proposed a novel image decomposition model and evaluated its performance against state-of-the-art works in image decomposition neural networks. Through numerical evaluation metrics, we showed that the proposed model performs consistently well with different datasets. We showed the robustness of the algorithm in a variety of scenes using visual examples. Though the proposed model outperforms existing works in terms of decomposition and reconstruction, there is room for improvement in relation to the color leakage problem in shading map.

6. REFERENCES

[1] F. Lv, Y. Li, and F. Lu, “Attention guided low-light image enhancement with a large scale low-light simulation dataset,” *arXiv preprint arXiv:1908.00682*, 2019.

[2] E. H. Land, “The retinex theory of color vision,” *Scientific American*, vol. 237, no. 6, pp. 108–129, 1977.

[3] C. Guo, C. Li, J. Guo, C. C. Loy, J. Hou, S. Kwong, and R. Cong, “Zero-reference deep curve estimation for low-light image enhancement,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 1780–1789.

[4] H. Weligampola, G. Jayatilaka, S. Sritharan, R. Godaliyadda, P. Ekanayaka, R. Ragel, and V. Herath, “A retinex based gan pipeline to utilize paired and unpaired datasets for enhancing low light images,” in *2020 Moratuwa Engineering Research Conference (MERCon)*. IEEE, 2020, pp. 224–229.

[5] B. T. Phong, “Illumination for computer generated pictures,” *Communications of the ACM*, vol. 18, no. 6, pp. 311–317, 1975.

[6] A. Watt and M. Watt, “Advanced animation and rendering techniques: Theory and practice,” 1992.

[7] M. Janner, J. Wu, T. D. Kulkarni, I. Yildirim, and J. Tenenbaum, “Self-supervised intrinsic image decomposition,” in *Advances in Neural Information Processing Systems*, 2017, pp. 5936–5946.

[8] X. Fu, D. Zeng, Y. Huang, X. Zhang, and X. Ding, “A weighted variational model for simultaneous reflectance and illumination estimation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2782–2790.

[9] B. Cai, X. Xu, K. Guo, K. Jia, B. Hu, and D. Tao, “A joint intrinsic-extrinsic prior model for retinex,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 4000–4009.

[10] J. Xu, Y. Hou, D. Ren, L. Liu, F. Zhu, M. Yu, H. Wang, and L. Shao, “Star: A structure and texture aware retinex model,” *IEEE Transactions on Image Processing*, vol. 29, pp. 5022–5037, 2020.

[11] S. Bell, K. Bala, and N. Snavely, “Intrinsic images in the wild,” *ACM Trans. on Graphics (SIGGRAPH)*, vol. 33, no. 4, 2014.

[12] L. Lettry, K. Vanhoey, and L. Van Gool, “Unsupervised deep single-image intrinsic decomposition using illumination-varying image sequences,” in *Computer Graphics Forum*. Wiley Online Library, 2018, vol. 37, pp. 409–419.

[13] Z. Li and N. Snavely, “CGIntrinsics: Better Intrinsic Image Decomposition through Physically-Based Rendering,” in *European Conference on Computer Vision (ECCV)*, 2018.

[14] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu, “Deep retinex decomposition for low-light enhancement,” in *British Machine Vision Conference (BMVC)*, 2018.

[15] A. S. Baslamisli, Y. Liu, S. Karaoglu, and T. Gevers, “Physics-based Shading Reconstruction for Intrinsic Image Decomposition,” *arXiv preprint arXiv:2009.01540*, 2020.

[16] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al., “Tensorflow: A system for large-scale machine learning,” in *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)*, 2016, pp. 265–283.

[17] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.

[18] A. Mittal, R. Soundararajan, and A. C. Bovik, “Making a “completely blind” image quality analyzer,” *IEEE Signal processing letters*, vol. 20, no. 3, pp. 209–212, 2012.

[19] A. Hore and D. Ziou, “Image quality metrics: PSNR vs. SSIM,” in *2010 20th international conference on pattern recognition*. IEEE, 2010, pp. 2366–2369.