TOWARDS HIGH-QUALITY INTRINSIC IMAGES IN THE WILD

Gang Fu[1], *Qing Zhang*[2], *Chunxia Xiao*[1]

[1]School of Computer, Wuhan University, Wuhan 430072, China
[2]School of Data and Computer Science, Sun Yat-sen University, Guangzhou 510006, China.
xyzgfu@gmail.com, zhangqing.whu.cs@gmail.com, cxxiao@whu.edu.cn

ABSTRACT

We address the intrinsic image decomposition problem for separating an image into its intrinsic images, *i.e*., a reflectance layer and a shading layer. Although this problem has been studied for decades, it remains a significant challenge, particularly for real-world images. In this paper, we present a novel method for estimating high-quality intrinsic images for real-world images. Our method is built upon two observations on real-world images: (*i*) reflectance is generally sparse and there are limited number of reflectance values in an image; (*ii*) shading usually has locally smooth transition. Based on the two observations, we formulate the decomposition problem into an optimization framework, where we encourage the reflectance sparseness by globally confining the number of reflectance discontinuities among neighboring pixels using an L_0 norm, and utilize a total variation for maintaining locally smooth shading. We employ two benchmark datasets and perform various experiments to evaluate our method. Experimental results show that our method outperforms state-of-the-art methods, both qualitatively and quantitatively.

***Index Terms*—** intrinsic image decomposition, real-world images, reflectance, shading

1. INTRODUCTION

Intrinsic image decomposition aims to separate an image I into its reflectance component R and shading component S as $I = R \times S$, with $\times$ denoting pixelwise multiplication. Since each component of the intrinsic images represents a meaningful physical property of the scene, many multimedia, computer vision and computer graphics tasks would benefit from such a decomposition, *e.g*., light information estimation [1], image-based resurfacing, relighting, recoloring, re-texturing, and 3D object compositing. Hence, intrinsic image decomposition has long been a fundamental problem in the research community.

This work was partly supported by The National Key Research and Development Program of China (2017YFB1002600), the NSFC (No. 61672390), Wuhan Science and Technology Plan Project (No. 2017010201010109), and Key Technological Innovation Projects of Hubei Province (2018AAA062). Chunxia Xiao is the corresponding author.

Intrinsic image decomposition remains a challenging and severely ill-posed problem because there exists twice as many unknowns (reflectance and shading) for each pixel. Although there has been significant progress on this problem, existing methods may fail to produce high-quality intrinsic images for real-world scene-level images (namely images in the wild), since this kind of image often contains a wide range of shapes and materials, and complex illumination.

Previous method have made considerable research efforts for estimating intrinsic images of real-world scenes. These methods can be roughly classified into two categories: optimization-based methods [2, 3] and learning-based methods [4]. However, as demonstrated in Fig. 1, they may fail to produce satisfactory results. For instance, Bell *et al*. [2] produce overly blurred structures in the reflectance layer, while Fu *et al*. [3] wrongly identify the shading variations on the ceiling as reflectance changes. The reflectance component generated by Li and Snavely [4] is dusty and unrealistic.

In this paper, we present a novel method to estimate high-quality intrinsic images for real-world images. The proposed method is built upon two observations, *i.e*., the *reflectance sparseness* and the *shading smoothness*. Based on the two observations, we formulate the intrinsic image decomposition as an energy minimization. Particularly, we encourage reflectance sparseness by globally confining the number of discontinuities within the reflectance component using an L_0 norm. Moreover, we enforce the shading smoothness by a total variation regularizer to alleviate the texture-copy problem. We solve the non-convex energy minimization by adopting the alternating direction of multipliers (ADMM) technique. Experimental results and comparisons to state-of-the-art methods validate the effectiveness of our method.

The major contributions of this work are as follows:

- We present a novel intrinsic image decomposition algorithm for real-world images, which incorporates an L_0 reflectance sparseness regularizer and total variation based shading smoothness regularizer.
- We derive a new procedure based on the alternating direction of multipliers (ADMM) technique for optimizing the proposed intrinsic decomposition model.
- We evaluate our method on two benchmark datasets

978-1-5386-1737-3/18/$31.00

(a) Input (b) Bell *et al*. [2] (c) Fu *et al*. [3] (d) Li and Snavely [4] (e) Ours

Fig. 1: Comparison between our proposed method with the state-of-the-art methods. **Please zoom in to compare results**.

and compare it with several state-of-the-art methods. Results show that our method outperforms previous methods. Moreover, we demonstrate that our method is also applicable to low-light image enhancement.

2. RELATED WORK

Currently, there is an immense literature on intrinsic image decomposition. In this section, we focus on discussing most related works rather than trying to be exhaustive.

Optimization-based methods. Due to the ill-posed nature of the intrinsic image decomposition, a majority of the approaches choose to integrate priors on reflectance and shading, and formulate the decomposition problem into an optimization framework. For instance, the pioneer Retinex algorithm [5] assumes large image gradients correspond to changes in reflectance, while smaller gradients are derived from shading discontinuities. Later, various priors have been proposed to guide the decomposition [6, 2]. However, for images of real-world scenes, such hand-crafted priors are often violated and would yield unsatisfactory results. Some recent approaches propose to use the surface normals from RGB-D cameras [7]. However, such methods assume high-quality depth maps are available, which limits their applicability to real-world consumer photographs.

Learning-based methods. Learning-based methods have recently attracted a large amount of attention, since they do not depend on the hand-crafted priors and can work automatically without tedious parameter tuning. Barron and Malik [8] learn a shape-from-shading model by employing priors on reflectance, shape and shading. This method works well for images of objects (such as images in the MIT intrinsic images dataset [9]), but does not generalize well to real-world scenes. Recently, deep-learning based methods have been widely deployed to predict reflectance and shading by training models on the synthetic Sintel [10], MIT dataset [9] or the synthetic ShapeNet datasets [11]. However, Sintel and ShapeNet are synthetic datasets, while the MIT dataset only contain 20 object-centric images captured under unrealistic scenes, making the networks trained on them may not suitable for real-world scenes.

Two recent datasets, IIW [2] and SAW [12], that consist of sparse, crowed-sourced reflectance and shading annotations on real-world scenes have greatly spurred advances on this problem. Using the datasets, several methods [13, 12] train CNN-based classifiers and use the classifier outputs as priors for intrinsic images. These methods report impressive results on the two datasets. However, since the annotations are sparse and the datasets fail to depict the diversity of the real-world scenes, they usually generalize poor to images that do not appear in the datasets.

3. PROPOSED METHOD

In this section, we describe our novel method for estimating intrinsic images. We first recall the image formation model, and then elaborate the proposed decomposition model. Finally, we introduce an efficient ADMM based solver for the non-convex optimization problem defined by our model, and the implementation and parameter setting details.

3.1. Image Formation Model

Let I denote an input image normalized to [0,1], and R and S denote the reflectance and shading components. By assuming the scene of the image I is Lambertian, we can write the intrinsic image decomposition problem as:

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

where S is a single-channel grayscale image because we assume the real-world scenes are lit by white light as most previous methods do, while R is multi-channel RGB image. Akin to [3], we also adopt the constraint $I \leq S$. Our goal is to estimate S and R from I. It is worth mentioning that our method can be directly extended to process images with color environment light by modeling S to be multi-channel.

3.2. Intrinsic Decomposition Model

We formulate the intrinsic image decomposition problem as the minimization of the following objective function:

$$\arg\min_{R,S} \mathcal{F}(R,S) = f_d + \lambda_r f_r + \lambda_s f_s + \lambda_a f_a \; s.t. \; I \leq S\,, \quad (2)$$

where f_d, f_r, f_s and f_a are different energy terms. λ_r, λ_s, and λ_a are all positive balancing weights. Below we describe each term in Eqn. (2) in detail.

Data fidelity term. According to Eqn. (1), valid intrinsic decomposition should be able to reproduce the input image by recombining the reflectance and shading with pixelwise multiplication. Hence, we define the following L_2 error metric to measure the reconstruction error:

$$f_d = ||I - R \times S||_2^2, \quad (3)$$

where $||\,||_p$ denotes the p-norm operator.

Reflectance sparseness term. Our one key observation of the real-world images is that the reflectance is generally sparse, *i.e.*, piecewise constant. Based on this observation, we encourage the reflectance sparseness by globally confining the number of reflectance discontinuities using a L_0 norm, which is defined as

$$f_r = ||\nabla R||_0, \quad (4)$$

where ∇ is the gradient operator. Intuitively, this L_0 sparsity term forces small reflectance discontinuities or noises to be zeros, while preserving the prominent structure of reflectance. This property benefits obtaining piecewise constant reflectance that is consistent to our observation, as demonstrated in Fig. 2(d). Previous methods [2, 3] also adopt the reflectance sparsity priors. In contrast, we advance them in two aspects. First, they basically employ L_2 [2] or L_1 [3] norm for sparsity, which make them not effective enough to produce piecewise constant reflectance, as shown in Fig. 2(b) and (c). Second, unlike them, our method is less sensitive to noise and complex shading variations.

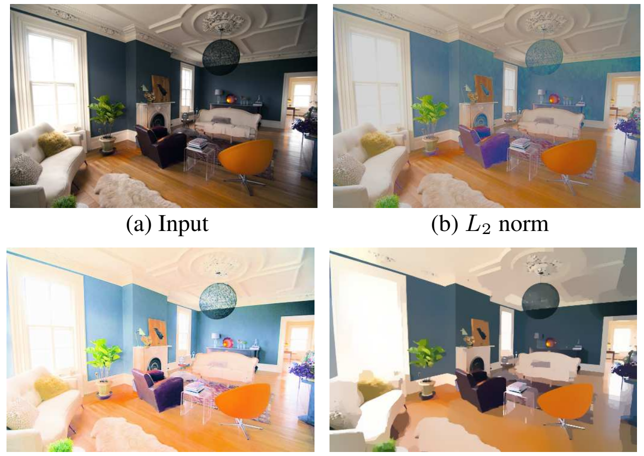
Shading smoothness term. Our another observation is that shading tends to vary smoothly across smooth surface. This observation on shading helps to determine the reflectance for textured object, since it is common knowledge that textures in real-world images are more likely to be incurred by reflectance instead of shading variations. Hence, we include a total variation based shading smoothness term:

$$f_s = ||\nabla S||_2^2. \quad (5)$$

As demonstrated in Fig. 3, this term benefits addressing the texture copy problem that resides in most existing methods.

Absolute scale term. Similar to [2], besides the shading smoothness term, we also impose a absolute scale constraint for shading, which is expressed as

$$f_a = ||S - \overline{S}||_2^2, \quad (6)$$



(a) Input (b) L_2 norm (c) L_1 norm (d) Ours (L_0 norm)

Fig. 2: Comparison between our proposed L_0 norm with L_2 and L_1 norm in encouraging reflectance sparsity. (b)-(d) are the estimated reflectance layers derived from different sparsity pursuit norms (L_2, L_1 and L_0).

(a) Input (b) Without f_s (c) With f_s

Fig. 3: Comparison of shading produced by our method without and with the shading smoothness term. We can see that, without the term, there are erroneous texture residuals in the shading layer, while adding the term removes the issue.

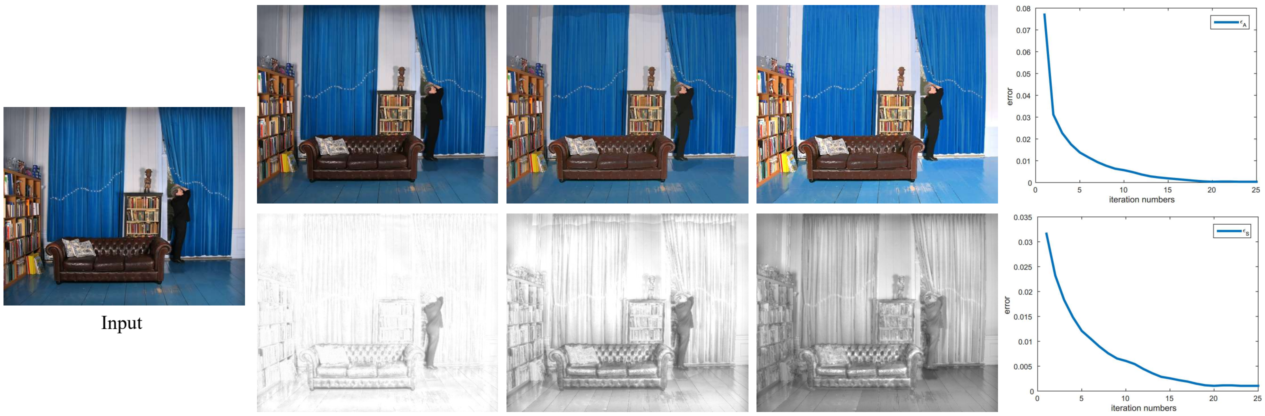
where $\overline{S}$ is a positive constant. In our experiments, we empirically set $\overline{S} = 0.5$. Adopting this scale constraint has two advantages. First, it helps avoid the ambiguity issue from $I = R \times S$, since for each decomposition we can enlarge S by any value and reduce R by some value to get a different valid decomposition. Second, it ensures that there are as few extreme values of shading as possible.

3.3. Model Solver

The objective function in Eqn. (2) is non-convex due to the L_0 norm regularization. We adopt the alternating direction method of multipliers (ADMM) technique to solve the optimization problem model by dividing the intractable problem into several tractable subproblems. Due to the space limit, we only give brief introduction. For more detailed description of the model solver, please refer to the supplementary material.

To apply ADMM, we introduce an auxiliary variable $G = \nabla R$ and an error variable X, and rewrite the objective function in Eqn. (2) as the following equivalent form:

$$\begin{aligned} &\arg\min_{R,S,G,X} ||I - R \times S||_2^2 + \lambda_s ||\nabla S||^2 + \lambda_a ||S - \overline{S}||_2^2 \\ &+ \lambda_r \{||G||_0 + \mu ||\nabla R - G + X||_2^2\}, \;\; s.t. \;\; I \leq S \end{aligned} \quad (7)$$



Input

Iteration 1 Iteration 10 Iteration 20 Convergence curve

Fig. 4: Results from different iterations and the convergence curves for reflectance (top) and shading (bottom).

where $\gamma > 0$ is the penalty parameter.

According to ADMM theory, the objective function in Eqn. (7) can be further divided into three subproblems with respect to G, R and S. By iteratively solving these subproblems in an alternating order until convergence, we can obtain solution to Eqn. (7). In particular, for the k-th iteration, the subproblems are as follows.

Subproblem 1: we solve for G^k by the minimization below:

$$\arg\min_{G} ||G||_0 + \mu^{k-1}||\nabla R^{k-1} - G + X^{k-1}||_2^2. \quad (8)$$

As introduced in [14], the above problem can be solved in an element-wise manner. Formally, for a pixel p, the solution is

$$G_p^k = \begin{cases} 0, & (Y_j^{k-1})^2 \leq \frac{1}{\mu^{k-1}} \\ Y_p^k, & \text{otherwise} \end{cases}, \quad (9)$$

where $Y^{k-1} = X^{k-1} - \nabla R^{k-1}$. Note that we initialize $X^0 = 0$ and $R^0 = I$, and assume X^{k-1} and R^{k-1} are seen as constants estimated from the previous iteration.

Subproblem 2: we can obtain R^k by solving the following minimization:

$$\arg\min_{R} \left|\left| R - \frac{I}{S^{k-1}} \right|\right|_2^2 + \lambda_r \mu^{k-1}||\nabla R - G^k + X^{k-1}||_2^2, \quad (10)$$

which is a quadratic function with closed-form solution. Here we adopt the Fast Fourier Transformation (FFT) to speed up the optimization. Specifically, R^k is updated by

$$R^k = \mathscr{F}^{-1}\left(\frac{I/(S^{k-1}+\sigma) + \lambda_r \mu^{k-1}\Phi}{\mathscr{F}(U) + \lambda_r \mu^{k-1} R^{k-1}(\sum_{d \in x,y} \mathscr{F}^*(\nabla_d) \cdot \mathscr{F}(\nabla_d))}\right), \quad (11)$$

where $\Phi = \mathscr{F}^*(\nabla_x) \cdot \mathscr{F}(G_x^k - X_x^{k-1}) + \mathscr{F}^*(\nabla_y) \cdot \mathscr{F}(G_y^k - X_y^{k-1})$, $\mathscr{F}$ is the FFT operator, $\mathscr{F}^*$ is the conjugate transpose, $\mathscr{F}^{-1}$ is the inverse FFT operator. ∇_x and ∇_y are first order derivative operator along horizontal and vertical direction, respectively. U is the identity matrix. σ is a small constant (typically 0.0001) that avoids division by zero. In this fashion, the matrix inversion operation is avoided since the derivative operator is diagonalized after FFT.

$\mu^k = 2\mu^{k-1}$ is multiplied by 2 in each iteration, and the error variable X^k is updated by

$$\begin{cases} X_x^k = X_x^{k-1} + \nabla_x R^k - G_x^k, \\ X_y^k = X_y^{k-1} + \nabla_y R^k - G_y^k, \end{cases} \quad (12)$$

Note that G, R, X and μ are estimated iteratively until $\epsilon_R = (||R^k - R^{k-1}||/||R^{k-1}||) \leq \epsilon_1$.

Subproblem 3: we solve for S^k by minimizing

$$\arg\min_{S} \left|\left| S - \frac{I}{R^k} \right|\right|_2^2 + \lambda_s||\nabla S||_2^2 + \lambda_a||S - \overline{S}||_2^2. \quad (13)$$

Similarly, the solution for S^k can be written as

$$S^k = \mathscr{F}^{-1}\left(\frac{\lambda_a \overline{S} + I/(R^k+\sigma))}{\mathscr{F}(U) + \lambda_a + \lambda_s(\sum_{d \in x,y} \mathscr{F}^*(\nabla_d) \cdot \mathscr{F}(\nabla_d))}\right). \quad (14)$$

Note that due to $R \in (0, 1]$ and $I \leq S$, after S^k is updated we further correct it by: $S = \max(S, I)$. S is updated until $\epsilon_S = (||S^k - S^{k-1}||/S^{k-1}) \leq \epsilon_2$.

3.4. Implementation and Parameter Setting

We implement our algorithm using Matlab on a laptop with Intel i7-6700 3.40GHZ CPU and 8GB of RAM. We have experimentally found that $\lambda_r = 0.001$, $\lambda_s = 0.01$ and $\lambda_a = 0.06$ works well for most testing images. The initial value of μ is 1. ϵ_1 and ϵ_2 are fixed to 0.001 in all our experiments. When $\epsilon_R \leq \epsilon_1$ and $\epsilon_S \leq \epsilon_2$, we stop the ADMM based iterative optimization procedure. It takes about 2 minutes for our algorithm to process an 800×600 image. Fig. 4 shows how the results updated with the iterations increased, and the convergence curves for reflectance and shading.

Table 1: Quantitative comparison of different algorithm variants of our method on the IIW and SAW datasets in terms of WHDR and SAW, respectively.

Algorithm variant	WHDR (IIW)	AP (SAW)
Without f_r (L_0)	23.54%	95.23%
f_r (L_0) using L_1 instead	27.43%	92.23%
f_r (L_0) using L_2 instead	38.23%	91.55%
Without f_s	24.32%	93.12%
Without f_a	29.34%	94.52%
Ours (full method)	**19.20%**	**98.80%**

4. EXPERIMENTAL RESULTS

4.1. Datasets and Evaluation Metrics

Benchmark datasets. Two benchmark datasets are employed to evaluate the effectiveness of our approach. The first one is the IIW dataset [2], which contains over 5,000 images covering a wide range of scenes, and has been annotated with about 100 human judgements on reflectance per image as ground truth data. The second benchmark dataset is the SAW dataset [12]. It includes 6,677 images with shading judgements obtained via crowdsouring, along with shading annotations generated from RGB-D imagery.

Evaluation metrics. We employ two commonly-used metrics to quantitatively evaluate the performance of the proposed method. The first one is the WHDR metric [2] defined in the IIW dataset. The second metric we employed is AP [12] for the SAW dataset. In general, lower WHDR and higher AP indicates better performance.

4.2. Comparison with the State-of-the-art Methods

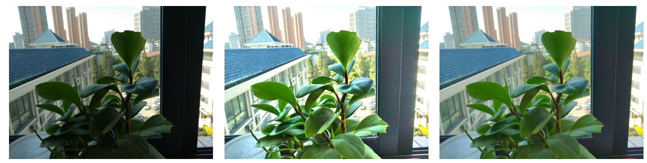
We compare our method with seven intrinsic image decomposition methods: Grosse *et al.* [9], Shen *et al.* [15], Garces *et al.* [16], Zhao *et al.* [6], Bell *et al.* [2], Fu *et al.* [3], and Li and Snavely [4]. Among them, the first six are optimization-based methods, while the last one is deep-learning-based method. For a fair comparison, the results of other methods are generated by implementations provided by the authors with the recommended parameter setting.

Table 2 reports the quantitative comparison results. We can see that our method outperforms the other compared methods in terms of the WHDR metric and AP metric for both the two benchmark datasets. Note that, although the state-of-the-art deep-learning-based method is directly trained on the IIW and SAW datasets, we also achieve comparable or even better results, and has better generalization capability.

We further provide visual comparison result in Fig. 7, where we can see that our method can effectively separate the reflectance and shading from the input image, while other compared methods either fail to distinguish shading variations from reflectance variations, or damage the structures in the reflectance layer and make the intrinsic images unrealistic.

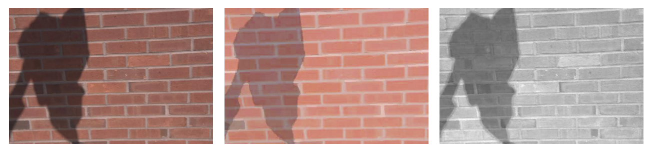
Table 2: Quantitative comparison between our method and the state-of-the-art methods on the IIW and SAW datasets.

Method	WHDR (IIW)	AP (SAW)
Retinex-color [9]	26.89%	91.93%
Shen *et al.* [15]	36.90%	87.23%
Garces *et al.* [16]	25.46%	96.89%
Zhao *et al.* [6]	23.20%	97.11%
Bell *et al.* [2]	20.64%	97.37%
Fu *et al.* [3]	36.43%	91.87%
Li and Snavely [4]	20.30%	97.90%
Ours	**19.20%**	**98.80%**



(a) Input (b) LIME [17] (c) Ours

Fig. 5: Application to low-light image enhancement.



(a) Input (b) Reflectance (c) Shading

Fig. 6: Failure case of our method.

4.3. Discussion

Ablation study. Besides the visual results in Figs. 2 and 3, we also quantitatively evaluate the effectiveness of the energy terms in our model. As shown in Table 1, replacing the L_0 reflectance sparseness term with L_2 and L_1 sparsity pursuit, or removing the shading smoothness term and the absolute scale term decrease the performance of our method for both the IIW and SAW datasets, which convincingly demonstrates the effectiveness of each energy term.

Application to low-light image enhancement. Our method is applicable to low-light image enhancement. Given a low-light image, we first estimate the illumination S by performing the intrinsic image decomposition, and adopt the gamma correction ($\gamma = 2.2$) to obtain a modified illumination S'. With the modified illumination, we then update the reflectance by $R' = I/S'$ and treat R' as the final enhanced image, as done in [17, 18]. Fig. 5 shows an example. We can see that, the LIME [17] overexposures the leaves and the sky, while our method produces more appealing result with moderate brightness, clear details, distinct contrasts and vivid colors.

Limitations and future work. The limitation of our method is that it may wrongly assign hard shadows that violate the Lambertian assumption to the reflectance layer. Fig. 6 shows an example. We can see that, the cast shadow led by shading

(a) Input

(b) Zhao *et al*. [6] (c) Bell *et al*. [2] (d) Fu *et al*. [3] (e) Li *et al*. [4] (f) Ours

Fig. 7: Visual comparison of intrinsic images produced by our method and some recent methods.

discontinuities also exists in the reflectance layer. In the future, we would like to adopt the scene semantic and the depth information to further remove this limitation. Second, our current implementation based on CPU is not fast enough for real-time applications. Hence, we leave it as future work to improve its performance by GPU computation.

5. CONCLUSION

We have presented a novel method for estimating high-quality intrinsic images for real-world scenes. Our method is built upon two key observations of the real-world scenes, namely the reflectance sparseness and the shading smoothness. Based on these two observations, we design an optimization framework. In particular, an L_0 norm is utilized to encourage the reflectance sparseness while preserving the prominent structures, and a total variation based shading smoothness term is adopted to avoid the texture-copy problem. We have conducted extensive experiments on the IIW and SAW datasets. Experimental results show that our method outperforms the state-of-the-art methods, both qualitatively and quantitatively.

6. REFERENCES

[1] L. Zhang, Q. Yan, Z. Liu, H. Zou, and C. Xiao, "Illumination decomposition for photograph with multiple light sources," *IEEE TIP*, vol. 26, no. 9, pp. 4114–4127, 2017.

[2] S. Bell, K. Bala, and N. Snavely, "Intrinsic images in the wild," *ACM TOG*, 2014.

[3] X. Fu, D. Zeng, Y. Huang, X. Zhang, and X. Ding, "A weighted variational model for simultaneous reflectance and illumination estimation," in *CVPR*, 2016.

[4] Z. Li and N. Snavely, "Learning intrinsic image decomposition from watching the world," in *CVPR*, 2018.

[5] E. H. Land. and J. J. McCann, "Lightness and retinex theory," *JOSA*, 1971.

[6] Q. Zhao, P. Tan, Q. Dai, L. Shen, E. Wu, and S. Lin, "A closed-form solution to retinex with nonlocal texture constraints," *IEEE TPAMI*, 2012.

[7] J. T. Barron. and J. Malik, "Intrinsic scene properties from a single rgb-d image," in *CVPR*, 2013.

[8] J. T. Barron and J. Malik, "Shape, illumination, and reflectance from shading," *IEEE TPAMI*, 2015.

[9] R. Grosse, M. K. Johnson, E. H. Adelson, and W. T. Freeman, "Ground truth dataset and baseline evaluations for intrinsic image algorithms," in *ICCV*, 2009.

[10] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black, "A naturalistic open source movie for optical flow evaluation," in *ECCV*, 2012.

[11] A. X. Chang, T. Funkhouser, L. Guibas, Hanrahan, Q. P., Huang, Z. Li, and J. Xiao, "Shapenet: an information-rich 3d model repository," *arXiv preprint arXiv:1512.03012*, 2015.

[12] B. Kovacs, S. Bell, N. Snavely, and K. Bala, "Shading annotations in the wild," in *CVPR*, 2017.

[13] T. Zhou, P. Krähenbühl, and A. A. Efros, "Learning data-driven reflectance priors for intrinsic image decomposition," in *ICCV*, 2015.

[14] L. Xu, C. Lu, Y. Xu, and J. Jia, "Image smoothing via l_0 gradient minimization," *ACM TOG*, 2011.

[15] J. Shen, Yang X., Y. Jia, and X. Li, "Intrinsic images using optimization," in *CVPR*, 2011.

[16] E. Garces, A. Munoz, J. Lopez-Moreno, and D. Gutierrez, "Intrinsic images by clustering," *CGF*, 2012.

[17] X. Guo, Y. Li, and H. Ling, "Lime: Low-light image enhancement via illumination map estimation," *IEEE TIP*, 2017.

[18] Q. Zhang, G. Yuan, C. Xiao, L. Zhu, and W. Zheng, "High-quality exposure correction of underexposed photos," in *ACM MM*, 2018.