Intrinsic Image Decomposition: Challenges and New Perspectives

Diclehan Ulucan, Oguzhan Ulucan and Marc Ebner
Institut für Mathematik und Informatik, Universität Greifswald
Walther-Rathenau-Straße 47, 17489 Greifswald, Germany
{diclehan.karakaya, oguzhan.ulucan, marc.ebner}@uni-greifswald.de

Keywords: Intrinsic image decomposition, performance evaluation, error metrics

Abstract: In the field of intrinsic image decomposition, alongside developing a robust algorithm for the ill-posed problem, it is also required to benchmark the method on a comprehensive dataset by using a suitable evaluation metric. However, there are certain limitations in existing evaluation metrics. In this study, two new evaluation strategies are proposed to analyze intrinsics according to their characteristics. The *ensemble of metrics* combines different perceptual quality metrics in scale-space, while the *imperceptible* ΔE *score* is the modified version of the classical ΔE metric. Intrinsic image decomposition studies that extract the reflectance and shading images are benchmarked on two datasets. Furthermore, an overview of the field of intrinsic image decomposition is provided and the challenges that have to be overcome are pointed out.

1 Introduction

We can easily perceive our surroundings by unconsciously making use of our abilities to estimate distances, discount the illuminant, and differentiate between colors (Zeki, 1993; Ulucan et al., 2022c). These abilities are difficult to mimic for machine systems. One way to enable an artificial system to carry out such tasks is to make use of intrinsic image decomposition. An image can be decomposed into a *"family of intrinsic characteristics"* (Barrow et al., 1978). Each element of this family is a low-level feature of the input scene and it is called an *intrinsic image*. Each intrinsic image allows us to extract distinct characteristics of a scene more efficiently (Ebner, 2007).

There are several problems in intrinsic image decomposition and one of the main challenges arises from its nature (Bonneel et al., 2017). Intrinsic image decomposition is a severely under-constrained problem, therefore most of the studies only consider the reflectance and shading features to simplify the problem. While the shading S can be described as the component demonstrating the interaction between the illumination and the surfaces, the reflectance R can be defined as the element providing the ratio between the total incident and total reflected illumination (Barrow et al., 1978; Shen et al., 2011). An image I at location (x, y) can be represented as follows;

$$I(x,y) = R(x,y) \cdot S(x,y). \qquad (1)$$

Another challenge in intrinsic image decomposition is the lack of a common evaluation benchmark. The utilized datasets have a tendency to meet the assumptions made in the proposed algorithms, hence objectively determining the best performing method is quite difficult (Bonneel et al., 2017). The fact that almost all datasets have distinct characteristics makes it also hard to assess intrinsic image decomposition methods in a robust manner. For instance, in the MIT Intrinsic Images dataset (Grosse et al., 2009) one object is placed in front of a black background without any strong shadow or color casts. In the Intrinsic Images in the Wild Dataset (Bell et al., 2014), the ground truth information is subjective. There exist also other large-scale datasets, however, some of them consist of images containing only a single 3D model rendered with an environmental map and in these datasets, the object can be easily segmented since it is placed in the foreground (Shi et al., 2017). It is also worth mentioning here that, there are some datasets, which are not specifically designed for intrinsic image decomposition but can be used in this field and others that contain very complex scenes (Li et al., 2021; Roberts et al., 2021). Based on these observations, we recently introduced a comprehensive intrinsic image decomposition benchmark called IID-NORD (Ulucan et al., 2022a), which contains ground truth information for 5 intrinsic images, namely, reflectance, shading, depth map, surface normals, and light vectors.

A further point considered as a challenge in intrinsic image decomposition studies is the absence of quality metrics reflecting the actual performance of the algorithms (Garces et al., 2022). Since existing image quality metrics focus on specific features that are important for the task at hand, it is difficult to find a metric performing robustly in a field such as intrinsic image decomposition, which requires the analysis of distinct images at once to make an overall ranking. Therefore, a metric, which analyzes distinct intrinsics by taking into account the individual features that each intrinsic holds and outputs a global quality score, would allow us to benchmark intrinsic image decomposition algorithms in a more accurate manner. On the other hand, since intrinsic images can be used in pipelines of different computer vision tasks, i.e. the reflectance can be adopted for image segmentation, metrics that are able to investigate an intrinsic individually are also needed.

Consequently, as the utilization of intrinsic image decomposition contributes to various computer vision tasks and computer graphics applications, it is essential to analyze the performances of existing intrinsic image decomposition algorithms in a robust manner to point out the shortcomings and strengths of the methods. This will also lead the path to design more efficient intrinsic image decomposition approaches. Thereupon, in this study, two new evaluation strategies, namely *"ensemble of metrics"*, and *imperceptible* ΔE *score*, which is a modified version of the ΔE metric, are introduced. Also, seven existing metrics are used to demonstrate the performance of intrinsic image decomposition methods. To the best of available knowledge although they can be beneficial for intrinsic image decomposition studies some of them have not been considered in this field yet. Furthermore, a subset of our recently introduced dataset IID-NORD is created, which contains only the reflectance and shading elements.

This paper is organized as follows. Section 2 presents intrinsic image decomposition algorithms. Section 3 introduces the new evaluation metrics. Section 4 discusses the experimental results. Section 5 gives a brief summary of the study.

2 Intrinsic Image Decomposition Methods

In the computer vision society, intrinsic image decomposition has been widely studied in the last decades, but the fundamental observations it is based on are dating back more than a thousand years to Alhazen, a famous scientist who left his substantial observations in optics as a legacy to the researchers in this field (Barrow et al., 1978; Barron and Malik, 2014).

The challenges it holds and the benefits it can provide made intrinsic image decomposition an attractive research field. Numerous algorithms based on various approaches and input requirements have been proposed in the last five decades. The intrinsic image decomposition methods may need multiple images taken under different lights, an input sequence where the light source is positioned at diverse locations in each image, a time-varying image stack, user scribbles, multiple images with distinct viewing conditions, depth information, different focal distances, or a single RGB input (Bonneel et al., 2017). An intrinsic image decomposition algorithm requiring only a single input image can be considered as more advantageous than methods relying on multiple images and different necessities. This observation relies on the fact that real-world single images are widely available and it is laborious to create image sequences in the appropriate format. Also, for tasks where intrinsic image decomposition is used as a pre-processing step, it is unlikely to have an input stack and inefficient to require user interaction. Based on these observations, in this study, algorithms relying on a single RGB image are considered for the experiments, since they reflect the requisites of many different applications. In this section, these algorithms are explained briefly.

The Retinex algorithm is one of the oldest intrinsic image decomposition studies (Land, 1964). The method is inspired from biological findings, which are based on Land's famous experiments. The algorithm relies on the observations that adjacent regions of distinct objects have sharp reflectance changes since the alteration between the intensities is large, and flat surfaces and shadows have smooth intensity differences. As a result, while the large gradient changes in an image are mostly due to changes in reflectance, small gradients are associated with the shading component. Later on, the Retinex algorithm is combined with a non-local reflectance constraint (Zhao et al., 2012). It is assumed that whenever two pixels have the same chromaticity texture vectors they also have the same reflectance value. Another intrinsic image decomposition algorithm is based on the assumption that considerably small patches in natural images should have similar reflectance values (Shen et al., 2011). Hence, the intrinsic image decomposition problem is solved by optimizing an energy function, where its constraints assign larger weights to the local neighboring pixels. In the SIRFS algorithm, intrinsic images are extracted from a masked image, which contains a single object (Barron and Malik, 2014). In this multiscale optimization based method prior information is taken to produce the intrinsics. In another intrinsic image decomposition study an unsupervised deep learning method, which trains on image pairs, is designed to recover the reflectance and shading information of a scene (Lettry et al., 2018). In the LR3M algorithm, which is a low-light enhancement method, intrinsics of an image are used to enhance the visual quality of the input while estimating a piece-wise smooth illumination and a noise-suppressed reflectance of the scene from a Retinex based model (Ren et al., 2020).

Apart from requiring a single input, these algorithms are selected for the evaluations, since they also have different characteristics, such as being based on optimization techniques, relying on neural networks, integrating intrinsic image decomposition into other image processing applications, and considering local spatial information, which increases the variety of analyses carried out in this study.

3 Evaluation Metrics

There are numerous intrinsic image decomposition algorithms but there is only one widely used objective error metric to benchmark the intrinsic image decomposition methods, namely the local mean squared error (LMSE), which is a modified version of the mean squared error (MSE) (Grosse et al., 2009). LMSE has difficulties in reflecting the actual performance of a method since the neighboring relationships of pixels are not considered (Bonneel et al., 2017; Garces et al., 2022). Therefore, the amount of available information in intrinsics and the LMSE score have a tendency to not coincide. Furthermore, when large regions having constant reflectance are decomposed correctly, usually very low LMSE is observed regardless of the remaining parts of the image (Bonneel et al., 2017).

Apart from LMSE, the peak signal-to-noise ratio (PSNR) and the structural similarity index (SSIM) are also used for evaluation. PSNR calculates the peak signal-to-noise ratio in decibels (dB) between the ground truth and processed images (Gonzalez and Woods, 2018). Since pixel-wise evaluations are conducted in PSNR, the neighboring relationships of pixels is neglected, and the scores do not necessarily represent the available information. On the other hand, SSIM takes into account the neighboring relationships of pixels and is inspired from the top-to-bottom assumption of the human visual system. SSIM investigates the structural similarity between the reference and processed images (Wang et al., 2004). The structure, contrast, and luminance components of images are regarded during the computation of the perceptual quality score. The image is evaluated patch-wise in SSIM, hence local spatial information is taken into account. SSIM scores are in the range $[0, 1]$, where a score closer to 1 refers to a better outcome. In order to avoid problems related to viewing conditions, later on, SSIM is modified and the multi-scale SSIM (MS-SSIM) is introduced (Wang et al., 2003). SSIM can be represented as follows (Wang et al., 2004);

$$SSIM = \frac{(2\mu_{I_1}\mu_{I_2} + C_1)(2\sigma_{I_1 I_2} + C_2)}{(\mu_{I_1}^2 + \mu_{I_2}^2 + C_1)(\sigma_{I_1}^2 + \sigma_{I_2}^2 + C_2)} \tag{2}$$

where, I_1 and I_2 are the ground truth and output, μ, σ, and σ^2 represent the mean, covariance, and variance, respectively, while C_1 and C_2 are small constants.

As it is pointed out in several studies (Gao et al., 2010; Ding, 2018; Zhu et al., 2021) evaluation strategies correlating with the human visual system have a tendency of achieving more reliable scores and the effectiveness of using these strategies for evaluating intrinsic image decomposition algorithms has already been proven (Chen and Koltun, 2013; Narihira et al., 2015). Additionally, it is well-known that carrying out computations using scale-space helps to avoid problems arising due to unknown viewing distance and display resolution. Thereupon, in this paper, to analyze distinct features of different intrinsics in a robust manner an *"ensemble of metrics"* is proposed, which utilizes different quality metrics in scale-space. In the ensemble of metrics (EM) three different evaluation strategies namely, SSIM, feature similarity index (FSIM) (Zhang et al., 2011), and visual information fidelity (VIF) (Sheikh and Bovik, 2006) are combined to benchmark intrinsic image decomposition studies. These metrics are selected, since they analyze features, which are important to the outcomes of intrinsic image decomposition algorithms, such as structure, contrast, luminance, color, and the amount of information coinciding between the ground truth and the estimated intrinsic image.

VIF (Sheikh and Bovik, 2006) aims at measuring how much of the information that can be extracted from the reference image can also be derived from the test image. The computation of VIF is carried out in the wavelet domain by making use of Gaussian scale mixtures C, which are a random field that can be presented as the product of two independent random fields. VIF can be computed as follows,

$$VIF = \frac{\sum_{k \in w} S_{\mathcal{RF}}(\vec{C}^{T,k}; \vec{F}^{T,k}|s^{T,k})}{\sum_{k \in w} S_{\mathcal{RF}}(\vec{C}^{T,k}; \vec{E}^{T,k}|s^{T,k})} \tag{3}$$

where, $S_{\mathcal{RF}}$ is the set of spatial locations for the random field, w represents the subbands of the image, $\vec{C}^{T,k}$ denotes T elements of C_k, $\vec{F}^{T,k}$ and $\vec{E}^{T,k}$ denote the T elements of the test image and reference images in one subband, respectively, and s^T is the model parameters of the associated image.

FSIM (Zhang et al., 2011) considers the local structures and contrast information of the images. FSIM is computed for grayscale images, but it has a straightforward extension for RGB images. FSIM is computed by using phase congruency (PC), which is contrast invariant, and gradient magnitude (GM). The PC component assumes that points having maximal phase in the frequency domain correspond to perceivable features, which correlates with the behavior of the human visual system while detecting significant features in images. Since the contrast information influences the human visual system during perception, GM is included during the formation of FSIM to take the contrast information of a scene into account. FSIM can be computed as follows,

$$FSIM = \frac{\sum_{x,y \in N}(F_{PC}(x,y) \cdot F_{GM}(x,y)) \cdot PC_{max}(x,y)}{\sum_{x,y \in N} PC_{max}(x,y)} \tag{4}$$

where, $F_{PC}(x,y)$ and $F_{GM}(x,y)$ are the PC and GM components of the image, respectively, $PC_{max}(x,y)$ represents the maximum PC value of the input images, and N is the number of pixels in the image. As VIF, FSIM also outputs scores in the range $[0,1]$, where scores closer to 1 represent better results.

In order to form the ensemble of metrics, first of all, the Gaussian and Laplacian pyramids of the input and estimation are computed. The number of scales is adaptively determined according to the image resolution. Both pyramids are utilized since they have different characteristics (Ebner et al., 2007). The Gaussian pyramid contains the low-frequency components of the image, thus most of the color information in the input is preserved in each scale, whereas the Laplacian pyramid behaves like a high-pass filter in which the high-frequency elements, i.e. fine details, of the images are maintained in every scale. The utilization of both of these pyramids leads to the consideration of distinct details at different scales, hence features of images can be analyzed in a more robust manner. Furthermore, analyzing the high- and low-frequency components in an image separately allows us to evaluate the outcomes of algorithms with metrics that are more suited to examine specific features appearing more explicitly in one pyramid than the other.

In the ensemble of metrics, SSIM, VIF, and FSIM are computed at each scale in both pyramids. While all metrics are utilized to evaluate the shading component, only SSIM and the colored FSIM (FSIMc) are considered for the reflectance element. VIF is discarded for reflectance since it is only computed for the luminance channel of images, i.e. the evaluation of the reflectance and shading components results in the same outcome. In the Gaussian pyramid, SSIM is calculated by taking into account all of the image features used in its standard computation, while in the Laplacian pyramid, only the structure and contrast are considered since the luminance component is irrelevant. On the other hand, all components of FSIM are regarded in both of the pyramids, since they are sensitive to the information in these pyramids. Lastly, VIF is computed in both pyramids, since high- and low-frequency components contain distinct information. After the scores at every scale in each pyramid are obtained, the scores of corresponding levels are linearly combined for each metric individually as follows,

$$P^R_{M'}(i) \quad = \quad \frac{G^R_{M'}(i) + L^R_{M'}(i)}{2} \tag{5}$$

$$P^S_M(i) \quad = \quad \frac{G^S_M(i) + L^S_M(i)}{2} \tag{6}$$

where, P is the average of the Gaussian and Laplacian pyramids, R and S are the reflectance and shading components, respectively, i represents the scale, $M' \in \{SSIM, FSIMc\}$ and $M \in \{SSIM, VIF, FSIM\}$.

Each P contains evaluation scores at various scales, hence these results have to be merged into one overall score for each metric. To fuse the scores, inspiration is taken from the experiments of forming the MS-SSIM metric (Wang et al., 2003). In the experiments of Wang *et al.*, which are based on human judgments, it is noticed that the human visual system gives different importance to the same error at distinct scales. Even when each image in a pyramid has the same error, the perceived quality changes in each scale. From the results of Wang's experiments, it can be deduced that the assigned importance is approximately Gaussian. Therefore, in EM, the scores at distinct scales are combined using a Gaussian-based weighting strategy to assign a different weight to each scale as follows,

$$P^R_{M'} \quad = \sum_i P^R_{M'}(i) \;\; e^{-\frac{i^2}{2\sigma^2}} \tag{7}$$

$$P^S_M \quad = \sum_i P^S_M(i) \;\; e^{-\frac{i^2}{2\sigma^2}} \tag{8}$$

where, σ depends on the number of scales and it is computed as $\sigma = (i-1)/5$.

Then, the scores for each intrinsic image are linearly combined as in the following,

$$EM_R \quad = \tfrac{1}{2} \sum_j P^R_{M'(j)} \tag{9}$$

$$EM_S \quad = \tfrac{1}{3} \sum_j P^S_{M(j)} \tag{10}$$

where, subscript j represents the j^{th} element of M' and M, and EM_R and EM_S are the ensemble of metrics scores for the reflectance and shading components, respectively. Lastly, the scores of reflectance and shading are averaged to obtain a global EM score. It should be noted here that the last level of the pyramids is ignored during FSIM computations in EM since FSIM also uses the scale-space during score computation, which causes ambiguities in the smallest level of the pyramids in EM.

As mentioned in Sec. 1, an evaluation strategy focusing on a single intrinsic image provides an effective analysis for tasks, where a particular intrinsic is of interest. Therefore, alongside EM, another metric, namely *imperceptible* ΔE *score*, is introduced in this study. This metric is considered only for the evaluation of the reflectance component since ΔE (CIEDE2000) (Luo et al., 2001; Sharma et al., 2005) focuses on the color difference of input images. ΔE is computed in CIELAB color space by analyzing the lightness, chroma, and hue components. While ΔE scores less than 1 are imperceptible, a score in the range $[1,4)$ may also be unnoticeable to observers (Ebner, 2007). Based on these findings, the conventional ΔE metric is modified. Since color information is a low-frequency component of images, the ΔE score is computed at each scale of the Gaussian pyramid. Then, the number of pixels having a ΔE score in the range $[0,4)$ is counted individually for every level. Subsequently, at each level, the number of pixels having an unnoticeable ΔE score is divided by the total number of pixels in the corresponding scale. Afterwards, all these ratios are weighted with a Gaussian function as in Eqn. 8 and summed to obtain the imperceptible ΔE score. As a result, a score closer to 1 indicates that the estimated reflectance image is approximating the ground truth, while scores closer to 0 show that significant observable differences between the ground truth and estimation are present. While the imperceptible ΔE metric is computed in scale-space due to its aforementioned advantages, it can also be calculated directly in the original scale of images. It is worth mentioning here that the imperceptible ΔE metric can also be useful for color constancy studies, where the standard ΔE metric is already being used as an evaluation strategy (Ebner, 2007; Ulucan et al., 2022b).

4 Experiments

In order to benchmark intrinsic image decomposition algorithms, a subset of our recent dataset IID-NORD is created in the open-source 3D graphics toolkit called OpenSceneGraph (www.openscenegraph.com). The same procedure with IID-NORD is followed, which is briefly explained in the following. The subset is called RS-NORD and it contains 1936 sRGB images along with their corresponding reflectance and shading ground truths. The images have a resolution of 1600×965 pixels. Each scene in RS-NORD contains a room with various objects. Different than IID-NORD, the textures are either created synthetically or captured using a mobile phone camera. The layout of the rooms and viewing angles are changed during rendering. A single point light source illuminating the scene with different lights is placed into the scene, and its location is repositioned for each rendering. Consequently, dynamic shadows (Wimmer et al., 2004) and distinct illumination conditions are obtained for the scenes. Also, the ambient light is turned on to make 20% of an object's color visible.

In this section the algorithms explained in Sec. 2, namely Retinex (Land, 1964; Grosse et al., 2009), Zhao (Zhao et al., 2012), Shen (Shen et al., 2011), SIRFS (Barron and Malik, 2014), Lettry (Lettry et al., 2018), and Ren (Ren et al., 2020) are benchmarked both on the MIT Intrinsic Images and the RS-NORD datasets by using an Intel i7 CPU @ 3.5 GHz Quad-Core 16 GB RAM machine. The implementations of the methods are taken from the official webpages of the authors. No optimization is carried out on the algorithms. Moreover, all the images are decomposed by a *baseline algorithm* (Bonneel et al., 2017), which is a simple approach that decomposes images without considering any important aspect of the intrinsic image decomposition problem. Any algorithm developed specifically for intrinsic image decomposition is desired to outperform the baseline method. In this approach, the chromaticity image (I_{ch}) is assumed to be the reflectance, and the square root of the direct average of channels (Y), i.e. grayscale illumination, is considered as the shading. I_{ch} and Y can be computed as follows,

$$I_{ch} = \left(\frac{R}{R+G+B}, \frac{G}{R+G+B}, \frac{B}{R+G+B}\right) \quad (11)$$

$$Y = \sqrt{\frac{R+G+B}{3}} \quad (12)$$

where, R, G and B are the color channels of the image.

In order to evaluate the algorithms, 9 different metrics namely, LMSE, PSNR, SSIM, MS-SSIM, FSIM, FSIMc, VIF, EM, and imperceptible ΔE score (ΔE_i) are used (Table 1). Note here that SIRFS is not evaluated on RS-NORD, since as mentioned in Sec. 2 this algorithm only takes input images with single objects.

Table 1: The statistical outcomes of algorithms. Best scores for each metric are highlighted. The last column provides the average execution time in seconds, where the run time of Ren is not provided since its code is binary and does not only output the intrinsics.

		LMSE	PSNR	SSIM	MS-SSIM	FSIM	FSIMc	VIF	EM	ΔE_i	Avg. time
MIT	Baseline	0.078	10.924	0.718	0.697	0.786	0.859	0.223	0.618	0.556	**0.031**
	Retinex	0.091	11.128	0.726	0.720	0.795	0.882	0.179	0.631	0.580	3.106
	Zhao	**0.036**	12.156	0.785	0.780	0.815	0.908	0.363	0.692	0.626	3.671
	Shen	0.062	13.753	0.698	0.756	0.725	0.864	0.298	0.709	0.639	38.097
	SIRFS	0.042	**13.812**	**0.803**	**0.797**	**0.833**	**0.911**	**0.377**	**0.724**	**0.714**	171.018
	Lettry	0.056	12.275	0.527	0.722	0.706	0.873	0.122	0.639	0.581	13.833
	Ren	0.079	9.233	0.703	0.713	0.816	0.869	0.176	0.605	0.528	–
RS-NORD	Baseline	0.093	10.030	0.599	0.636	0.776	0.611	**0.391**	0.622	0.014	**0.242**
	Retinex	0.101	10.581	**0.658**	0.688	**0.786**	0.677	0.386	**0.652**	0.061	53.634
	Zhao	0.102	6.257	0.305	0.633	0.772	0.606	0.117	0.441	0.039	40.219
	Shen	**0.068**	11.457	0.575	0.612	0.708	0.681	0.313	0.612	**0.079**	390.300
	Lettry	0.097	**12.093**	0.621	**0.710**	0.766	**0.702**	0.235	0.643	0.018	220.351
	Ren	0.094	9.562	0.505	0.670	0.738	0.653	0.121	0.585	0.024	–

As it can be seen from the results of the MIT Intrinsic Images dataset (Table 1), Zhao has the lowest LMSE, while SIRFS has the best scores in all the other metrics. This is also in accordance with the visual results in Fig 1. As aforementioned LMSE neglects local spatial cues, and may not coincide with the perceptually available information that is taken into account in metrics such as FSIM and FSIMc, which indicates that LMSE may not reflect the actual performance of algorithms. As demonstrated in Fig. 1, Zhao faces an obvious challenge in eliminating shadows and specularity from the reflectance image, while Shen is mostly able to handle these features, which also coincides with the ΔE_i and EM scores. As discussed in Sec. 3, evaluation strategies correlating with the human visual system generally output more accurate results. Hence, investigating the visual outcomes of the intrinsic image decomposition methods can help to understand what type of metrics provide a more reliable score. It can be argued that the proposed metrics are able to output scores that correlate with the actual available information in the intrinsics.

The intrinsic image decomposition algorithms face a challenge when benchmarked on a more complex dataset than the MIT Intrinsic Images dataset. Generally, for each metric, a different method produces the best score in RS-NORD. In terms of PSNR, MS-SSIM, and FSIMc, Lettry outperforms the other intrinsic image decomposition methods. However, the visual results demonstrate that Lettry outputs color-distorted intrinsics, which reduces its ΔE_i significantly, and affects its EM score. According to the visual outcomes in the RS-NORD dataset, overall, Retinex produces the closest intrinsics to the ground truth information, which can also be observed from its EM score. On the other hand, LMSE, which is designed for intrinsic image decomposition studies, indicates that Shen performs the best decomposition among others. However, as seen in Fig. 1, while Shen greatly preserves the color information in the reflectance element, which coincides with its ΔE_i score, it faces difficulty in extracting the shading information, which is reflected in its EM score. While ambiguities are present in both intrinsics, the reason LMSE highlights Shen as the best performing algorithm can be explained by the fact that the preservation of large areas of constant reflectance tends to result in low LMSE scores (Bonneel et al., 2017). As it can be deduced from these observations, during the evaluation of intrinsic image decomposition algorithms it is important to consider the different characteristics of each intrinsic image and weigh the outcomes in a balanced manner in order to output a reliable statistical score.

Additional results of algorithms together with their scores are provided in Fig 2. The proposed metrics are able to output results coinciding with the available information in the intrinsics. In cases, where the ΔE_i and EM scores do not coincide, it can be deduced that while one of the intrinsics is successfully estimated, there is an issue in the estimation of the other one.

When intrinsic image decomposition methods are tested on a complex dataset both the statistical and visual results show that a single approach in solving the ill-posed intrinsic image decomposition problem is not sufficient to handle image features such as strong shadow casts, highlights, and specularities, since each method is more responsive to different features. Therefore, an ensemble of intrinsic image decomposition methods, which consists of algorithms that can handle different features might be beneficial.

5 Conclusion

Intrinsic image decomposition is an extensively studied field, which holds many challenges. The under-constraint nature of intrinsic image decomposition is the main difficulty in this field. Even though simplifications are made during intrinsic image computations, the ill-posed structure of the problem remains. Other challenges in intrinsic image decomposition are the lack of a common benchmark and an evaluation metric that allow us to analyze the intrinsic image decomposition algorithms efficiently. In this study, two new evaluation techniques, namely ensemble of metrics and imperceptible ΔE, are proposed to present possible solutions and new perspectives to the field of intrinsic image decomposition. Moreover, it is aimed to provide a guide to future studies by giving an overview of the field and addressing the problems it holds.

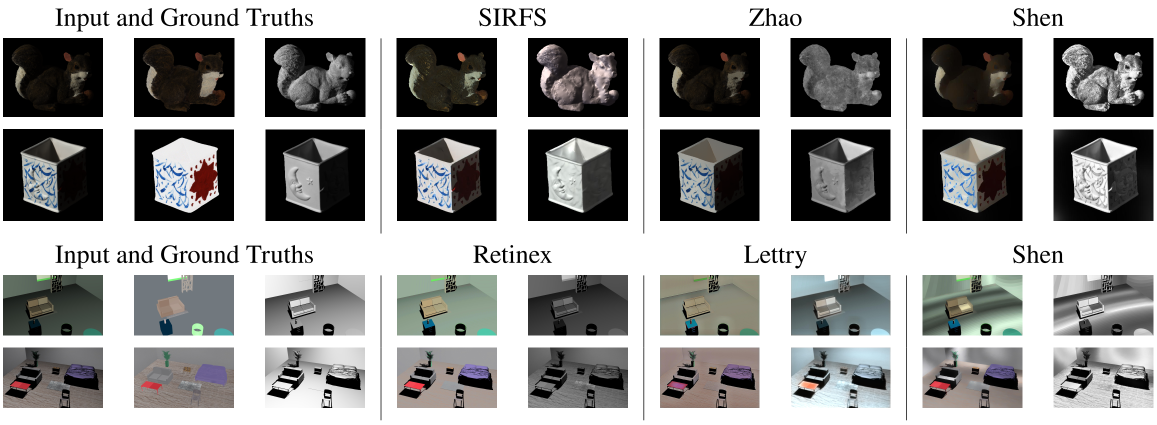

Figure 1: Visual comparison of algorithms. First two rows present scenes from the MIT Intrinsic Images dataset, and last two rows contain images from the RS-NORD dataset.

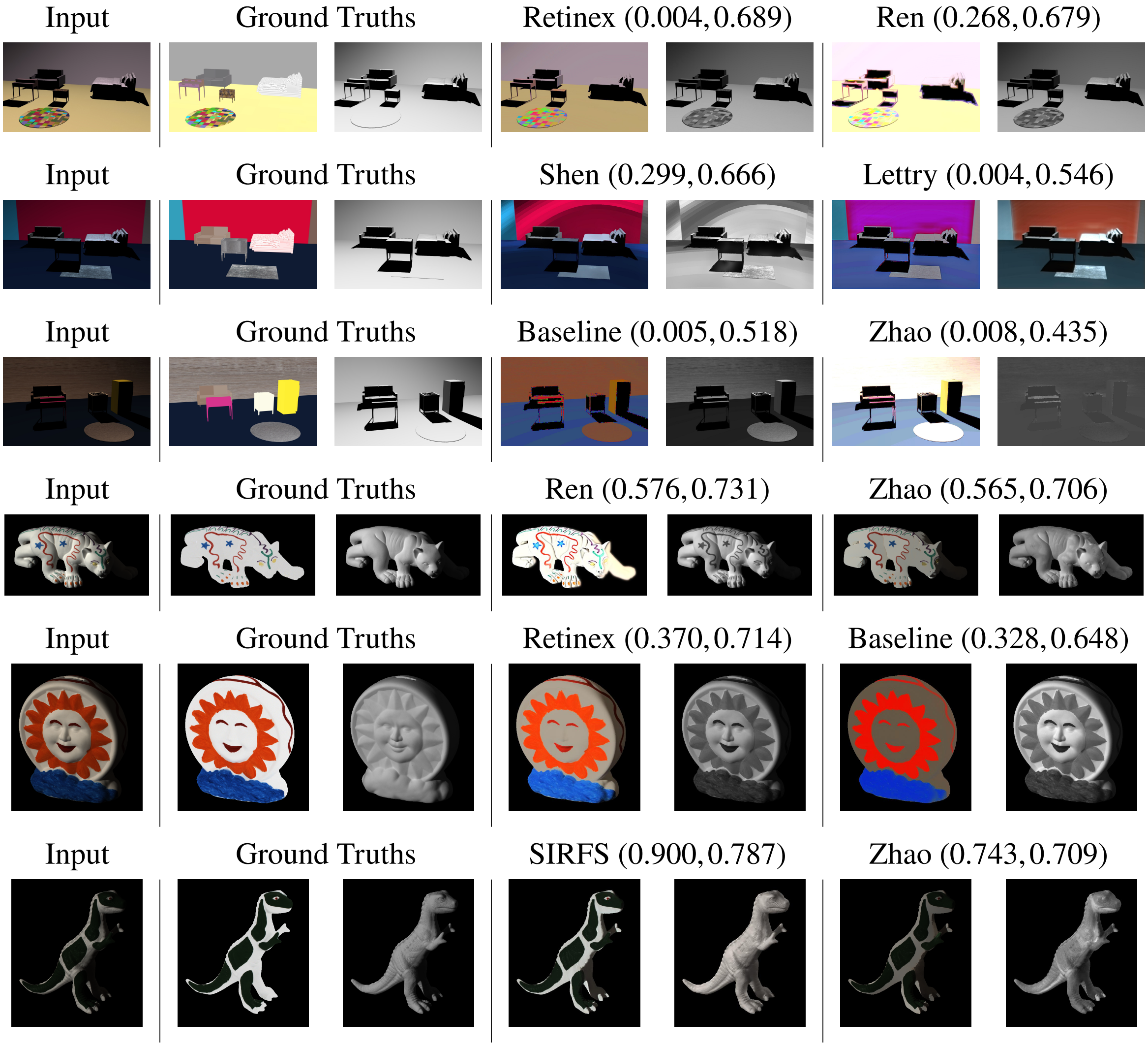

Figure 2: Comparison of the algorithms on both datasets (first three rows are from RS-NORD, and last three rows are from the MIT Intrinsic Images dataset). For each method firstly the ΔE_i score, then the EM score is given in parenthesis.

REFERENCES

Barron, J. T. and Malik, J. (2014). Shape, illumination, and reflectance from shading. *IEEE Trans. Pattern Anal. Mach. Intell.*, 37:1670–1687.

Barrow, H., Tenenbaum, J., Hanson, A., and Riseman, E. (1978). Recovering intrinsic scene characteristics. *Comput. Vision Syst.*, 2:2.

Bell, S., Bala, K., and Snavely, N. (2014). Intrinsic images in the wild. *ACM Trans. Graph.*, 33:1–12.

Bonneel, N., Kovacs, B., Paris, S., and Bala, K. (2017). Intrinsic decompositions for image editing. *Comput. Graph. Forum*, 36:593–609.

Chen, Q. and Koltun, V. (2013). A simple model for intrinsic image decomposition with depth cues. In *ICCV*, pages 241–248, Sydney, NSW, Australia. IEEE.

Ding, Y. (2018). Image quality assessment based on human visual system properties. *Vis. Qual. Assessment Natural Med. Image*, 37:63–106.

Ebner, M. (2007). *Color Constancy, 1st ed.* Wiley Publishing, ISBN: 0470058299.

Ebner, M., Tischler, G., and Albert, J. (2007). Integrating color constancy into JPEG2000. *IEEE Trans. Image Process.*, 16:2697–2706.

Gao, X., Lu, W., Tao, D., and Li, X. (2010). Image quality assessment and human visual system. In *Proc. Vis. Commun. Image Process.*, pages 316–325, Huangshan, China. SPIE.

Garces, E., Rodriguez-Pardo, C., Casas, D., and Lopez-Moreno, J. (2022). A survey on intrinsic images: Delving deep into lambert and beyond. *Int. J. Comput. Vision*, 130:836–868.

Gonzalez, R. C. and Woods, R. E. (2018). *Digital Image Processing, 3rd ed.* Pearson Prentice Hall.

Grosse, R., Johnson, M. K., Adelson, E. H., and Freeman, W. T. (2009). Ground truth dataset and baseline evaluations for intrinsic image algorithms. In *ICCV*, pages 2335–2342, Kyoto, Japan. IEEE.

Land, E. H. (1964). The retinex. *Amer. Scientist*, 52:247–264.

Lettry, L., Vanhoey, K., and Van Gool, L. (2018). Unsupervised deep single-image intrinsic decomposition using illumination-varying image sequences. *Comput. Graph. Forum*, 37:409–419.

Li, Z., Yu, T.-W., Sang, S., Wang, S., Song, M., Liu, Y., Yeh, Y.-Y., Zhu, R., Gundavarapu, N., Shi, J., Bi, S., Yu, H.-X., Xu, Z., Sunkavalli, K., Hasan, M., Ramamoorthi, R., and Chandraker, M. (2021). Openrooms: An open framework for photorealistic indoor scene datasets. In *CVPR*, pages 7190–7199, Nashville, TN, USA. IEEE.

Luo, M. R., Cui, G., and Rigg, B. (2001). The development of the CIE 2000 colour-difference formula: CIEDE2000. *Color Res. Appl.*, 26:340–350.

Narihira, T., Maire, M., and Yu, S. X. (2015). Direct intrinsics: Learning albedo-shading decomposition by convolutional regression. In *ICCV*, pages 2992–2992, Santiago, Chile. IEEE.

Ren, X., Yang, W., Cheng, W.-H., and Liu, J. (2020). LR3M: Robust low-light enhancement via low-rank regularized retinex model. *IEEE Trans. Image Process.*, 29:5862–5876.

Roberts, M., Ramapuram, J., Ranjan, A., Kumar, A., Bautista, M. A., Paczan, N., Webb, R., and Susskind, J. M. (2021). Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *ICCV*, pages 10912–10922, Montreal, QC, Canada. IEEE.

Sharma, G., Wu, W., and Dalal, E. N. (2005). The CIEDE2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations. *Color Res. Appl.*, 30:21–30.

Sheikh, H. R. and Bovik, A. C. (2006). Image information and visual quality. *IEEE Trans. Image Process.*, 15:430–444.

Shen, J., Yang, X., Jia, Y., and Li, X. (2011). Intrinsic images using optimization. In *CVPR*, pages 3481–3487, Colorado Springs, CO, USA. IEEE.

Shi, J., Dong, Y., Su, H., and Yu, S. X. (2017). Learning non-lambertian object intrinsics across shapenet categories. In *CVPR*, pages 1685–1694, Honolulu, HI, USA. IEEE.

Ulucan, D., Ulucan, O., and Ebner, M. (2022a). IID-NORD: A comprehensive intrinsic image decomposition dataset. In *ICIP*, pages 2831–2835, Bordeaux, France. IEEE.

Ulucan, O., Ulucan, D., and Ebner, M. (2022b). BIO-CC: Biologically inspired color constancy. In *BMVC*, London, UK. BMVA Press.

Ulucan, O., Ulucan, D., and Ebner, M. (2022c). Color constancy beyond standard illuminants. In *ICIP*, pages 2826–2830, Bordeaux, France. IEEE.

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 13:600–612.

Wang, Z., Simoncelli, E. P., and Bovik, A. C. (2003). Multiscale structural similarity for image quality assessment. In *Proc. Asilomar Conf. Signals Syst. Comput.*, pages 1398–1402, Pacific Grove, CA, USA. IEEE.

Wimmer, M., Scherzer, D., and Purgathofer, W. (2004). Light space perspective shadow maps. *Rendering Techn.*, 2004:143–151.

Zeki, S. (1993). *A Vision of the Brain*. Blackwell Science, ISBN: 0632030545.

Zhang, L., Zhang, L., Mou, X., and Zhang, D. (2011). FSIM: A feature similarity index for image quality assessment. *IEEE Trans. Image Process.*, 20:2378–2386.

Zhao, Q., Tan, P., Dai, Q., Shen, L., Wu, E., and Lin, S. (2012). A closed-form solution to retinex with nonlocal texture constraints. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34:1437–1444.

Zhu, W.-H., Sun, W., Min, X.-K., Zhai, G.-T., and Yang, X.-K. (2021). Structured computational modeling of human visual system for no-reference image quality assessment. *Int. J. Automat. Comput.*, 18:204–218.