Semantic Priors for Intrinsic Image Decomposition

Saurabh Saini
saurabh.saini@research.iiit.ac.in
P. J. Narayanan
pjn@iiit.ac.in

CVIT, KCIS,
International Institute of Information Technology - Hyderabad,
India

Abstract

Intrinsic Image Decomposition (IID) is a challenging and interesting computer vision problem with various applications in several fields. We present novel semantic priors and an integrated approach for single image IID that involves analyzing image at three hierarchical context levels. *Local* context priors capture scene properties at each pixel within a small neighborhood. *Mid-level* context priors encode object level semantics. *Global* context priors establish correspondences at the scene level. Our semantic priors are designed on both fixed and flexible regions, using selective search method and Convolutional Neural Network features. Experiments and analysis of our method indicate the utility of our weak semantic priors and structured hierarchical analysis in an IID framework. We compare our method with the current state-of-the-art and show results with lesser artifacts. Finally, we highlight that proper choice and encoding of prior knowledge can produce competitive results compared to end-to-end deep learning IID methods, signifying the importance of such priors. We believe that the insights and techniques presented in this paper would be useful in the future IID research.

1 Introduction and Related Work

Intrinsic image decomposition (IID) is a classic computer vision problem for splitting a given image (I) into two underlying components: $I = R \cdot S$ where R (reflectance) captures the color properties of the objects and S (shading) represents direct and indirect lighting in the scene. IID is useful in several computer vision and image editing applications like image colorization [37], shadow removal [29], image re-texturing [12], scene relighting [17] *etc*. IID is an ill-defined and under-constrained problem [6]. Moreover, IID solutions are also inherently ambiguous as there can be multiple valid reflectance and shading decompositions differing by a scalar multiplicative factor [9]. Hence several IID methods depend on auxiliary scene data in the form of depth [9, 14, 25], user scribbles [10], optical flow [28], multiple views [51], multiple illuminations [48], focal stacks [42], photo collections [30, 37], *etc*. The common idea behind such methods is to approximate textural and shape similarities using the auxiliary information but having the necessity to acquire additional data as a major drawback of such methods.

A second category of IID methods work directly on single images. These methods work under several assumptions and priors as it is hard to gather sufficient information about geometry, material property and illumination of the scene from a single image. Many such

methods work on simple images containing a single object with no background [2, 3, 5]. Other methods which work on natural scenes utilize priors like *Retinex* [32], reflectance sparsity [20, 45], long vs. short tailed gradient distribution [34], spatio-chromatic clustering [19], *etc*. These methods encode interesting insights but are limited when generalizing to 'wild' cases with varying lighting and complex textures. Results vary based on how much significance is given to a prior and the type of optimization framework. Moreover some of these priors have competing goals *e.g*. smoothness prior on shading removes texture details from S as opposed to reflectance sparsity assumption which simplifies color details in R. Recent methods try to solve this issue by sequentially employing two separate optimizations for shading and reflectance [6, 7, 51]. Based on this insight, our algorithm combines these two types of optimizations in a single integrated algorithm by alternating between two competing formulations: smoothness for shading and sparsity for reflectance.

A major challenge associated with IID research is lack of diverse large datasets and proper evaluation metrics [9]. This arises mainly due to subjective nature of the problem and difficulty in collecting dense annotations. MIT intrinsic images dataset introduced by Grosse et al. [23] is limited to a handful of single object images on a black background. Similarly As-Realistic-As-Possible (ARAP) dataset by Bonneel et al. [9] tries to capture complexity of natural scenes but is also not large enough for supervised training. Synthetic datasets like Sintel [11] provide dense annotation but lack sufficient diversity and complexity compared to the natural scenes. This limits the utility of such datasets in learning based approaches which aim to work on complete scenes under unrestricted illumination and material property settings. Yet another challenge in IID research is lack of proper evaluation metric which reflects both quantitative and qualitative performance. Local Mean Square Error (LMSE) and Structural Similarity Index Metric (SSIM) are used for synthetic scenes [14, 25] but require dense ground truth annotations. Bell et al. [6] provide a large manually annotated dataset called Intrinsic Images in the Wild (IIW) with sparse relative annotations. Their performance evaluation metric Weighted Human Disagreement Rate (WHDR) gives relative error rate using these sparse annotations. Some IID methods use supervised learning to solve related sub-problems using gradient classifiers [46], Bayesian graphical models [13] and deep neural networks [27, 40, 44, 51]. In [51] and [52] authors learn Convolutional Neural Network (CNN) priors using sparse IIW annotations which they propagate to other pixels using a dense Conditional Random Field (CRF) or flood fill the superpixels. Yet another approach is to use dense ground truth from synthetic scenes like [11]. Such approaches either use the underlying depth information [40] or use previously proposed RGBD based IID solutions to generate ground truth [27]. Synthetic datasets like Sintel do not represent true reflectance and shading of natural scenes as the dataset was not originally curated with the intention of IID benchmarking [25]. Due to limited data and significant domain shift, such end-to-end CNNs are prone to over-fitting and dataset bias [26, 47]. This inference was also highlighted by Nestmeyer and Gehler [41] who showed how a simple post processing using guided filtering could improve results from several deep learning IID solutions suggesting that such solutions are not able to capture the insights of the known IID priors effectively. Such issues concerning datasets and evaluation along with ill-defined nature of the problem, make IID a non-trivial problem to solve using deep learning. On the other hand, several older IID methods were unsupervised [10, 20, 48] and relied on intelligently designed priors. In our method we try to absorb the advantages of both the approaches. We employ an 'off-the-shelf' pre-trained deep neural network as a black-box to obtain generic features and introduce new context priors in an unsupervised optimization algorithm. CNNs have been widely used in computer vision and machine learning literature as black box feature

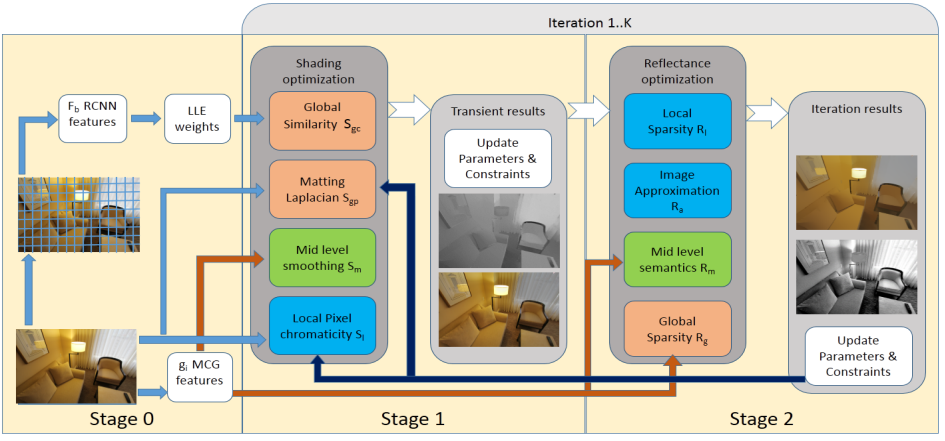


Figure 1: Block Diagram: Our method can be understood in three stages. After semantic features extraction (Stage 0), in each iteration our method alternates between the shading (Stage 1) and reflectance optimization (Stage 2) with energy terms computed for both the formulations computed at three context levels: local, mid-level and global.

extractor [16, 43, 49]. Donahue et al. [16] directly use pre-trained CNNs as a feature extractor and prove the generality and cross domain applicability of such features on varied tasks like scene recognition, fine-grained recognition and domain adaptation. Along similar lines, Sharif Razavian et al. [43] and Yosinski et al. [49] also use these features on increasingly different tasks and datasets, highlighting their task agnostic characteristics.

Our main contributions in this paper are: *(i)* We introduce a new technique for encoding scene semantic information for both fixed and flexible region definitions using CNN and selective search features. *(ii)* We present a new iterative integrated IID framework based on Split-Bregman iterations [22] using two competing IID formulations and generate results with lesser artifacts. *(iii)* We analyze scene at three context levels: *local context* where optimization weights are based on a small pixel neighborhood; *mid-level context* which tries to capture object level semantics and *global context* where various regions of the image are linked based on their shared characteristics at the scene level.

We perform experiments to analyze the effect of our new semantic priors at various context levels and illustrate the decompositions generated by our competing formulations over successive iterations. Finally, we present improved qualitative and competitive quantitative results with respect to the contemporary IID methods on challenging IIW dataset and 'wild' images from the Internet.

2 Method

Our method is as an iterative algorithm alternating between shading and reflectance formulations (Figure 1). Optimizing for reflectance sparsity alone leads to loss of textures in reflectance while focusing on shading smoothness does not account for reflectance sparsity. We tackle this adversarial nature of the two formulations by estimating IID in two separate stages for shading smoothness and reflectance sparsity. An iterative scheme has earlier been used by Bell et al. [6] and later adapted by Zhou et al. [51]. Our framework differs from them

as we present a single integrated algorithm without requiring additional steps for building a dense CRF or a separate optimization framework. We take inspiration from Bi et al. [7] who use Split-Bregman L_1-L_2 optimization method ([22]) for image flattening and adapt it to directly estimate IID. We show that this cleaner integrated approach leads to lesser artifacts in the results while maintaining good quantitative performance. In order to capture semantic scene information, we extract two different kinds of features for both fixed grids and flexible regions which are discussed below:

2.1 *Semantic Features*

RCNN features (f_b): We divide the input image I into B patches using a fixed non-overlapping grid of size 30×30. We pass these patches through Region-based Convolutional Neural Network (RCNN [21]) pre-trained on ImageNet dataset [15] and extract 4096 dimensional features f_b for each patch with $b \in \{1,2,\ldots B\}$ from the last fully connected layer ($fc7$) of the network. We assign this to the center pixel of the patch to obtain a sparse set of regional features for the image. Long et al. [38] show that such features, despite having weak label training over the entire scene and large receptive fields, encode fine correspondences between regions as in *SiftFlow* [36] and hence could be used even in tasks requiring precise localization like intraclass alignment and keypoints classification. We use f_b to estimate shape similarity and correspondences between image patches.

Selective search features (g_i): Selective search or detection proposals [24] give interesting image regions which have higher probability of containing an object. This improves object detection by avoiding sliding-window search. Hence selective search results could be used as an indicator of presence of an object ('objectness' [24]) in a given region. Compared to training a Conditional Random Field (CRF) for a finite number of classes like [7, 51] for dense pixel associations, selective search is class agnostic, has off-the-shelf implementations available and does not require separate training. We use Multiscale Combinatorial Grouping (MCG) [1] for capturing object semantics following the conclusions based on recall and detection quality from the survey of selective search techniques by Hosang et al. [24]. MCG generates dense binary region masks and scores for each detection proposal $c \in \{1,2,\ldots,P\}$ for a total of P proposals. We form a concatenated feature vector g_i of proposal masks weighed by proposal score at each pixel and normalize it using L_2 norm (See supplementary for sample masks and visualization).

2.2 Shading Formulation

Our shading formulation assumes monochromatic Lambertian illumination and piecewise constant reflectance [9] and is inspired from [25] which uses depth maps to define pixel neighbourhoods. We generalize their system for a single image by modifying the priors using RCNN and selective search features. The intermediate IID results as shading (σ) and reflectance (ρ), are estimated by minimizing the following energy function :

$$\Psi = \lambda_g S_g + \lambda_m S_m + \lambda_l S_l. \tag{1}$$

Here S_g, S_m and S_l are respectively global, mid-level and local shading priors and λ_g, λ_m and λ_l are the corresponding weights.

Global context (S_g)**:** Our global shading prior S_g is a combination of a sparse neighbourhood consistency term S_c and a weight propagation term S_p: $S_g = S_c + S_p$. In [25]

authors show that under the assumption of Lambertian model, shading at a point can be approximated using a weighted linear combination of surface normals where the weights are computed using Local Linear Embedding (LLE) in the neighbourhood $\mathcal{N}$. But unlike [25] we do not have depth information and therefore we approximate structural similarity using our pre-computed RCNN features f_b as:

$$S_c = \sum_b \big(\sigma_b - \sum_{a \in \mathcal{N}_b} w^c_{ab} \sigma_a\big)^2. \tag{2}$$

Here $\mathcal{N}_b$ represents the set of 10-nearest neighbours for patch b computed using f_b features and w^c are linear combination weights computed using the LLE representation of b over $\mathcal{N}_b$. These are sparse constraints as we assume the center pixel to be the representative of the entire patch and assign the constraint to it. In order to propagate these these constraints to the rest of the pixels, we do structure-aware weight propagation using a Laplacian matting matrix [33]. This approximates shading by an affine function over a base image in a small local window ($\mathcal{N}_{3\times3}$).

$$S_p = \sum_i \sum_{j \in \mathcal{N}_{3\times3}} w^p_{ij} \big(\sigma_i - \sigma_j\big)^2. \tag{3}$$

Here weights w^p are computed using the matting Laplacian [33] with reflectance result of the previous iteration as the base image. For the initial iteration, the base image for the laplacian is taken as Gaussian smoothened version of I. In [6] global constraints are propagated using a dense CRF whereas Zhou et al. [51] devised a Nyström approximation to integrate their proposed CNN reflectance prior for message passing during CRF inference. In comparison, Laplacian matting term has a closed form solution and is easy to compute [25].

Mid-level context (S_m): For mid-level prior we use selective search features g_i which encode object semantics. Similar to the weight propagation term S_p, we define this prior as:

$$S_m = \sum_i \sum_{j \in \mathcal{N}_{3\times3}} w^m_{ij} \big(\sigma_i - \sigma_j\big)^2 \tag{4}$$

where $w^m_{ij} = \exp\big(-\frac{(1-\langle g_i, g_j\rangle)^2}{t_m^2}\big)$ which penalizes dissimilar g_i and g_j. This captures the intuition that in a local neighbourhood if two pixels are predicted to belong to a common object proposal, then they should have similar shading. This causes shading smoothness within each detection proposals and preserves texture in the reflectance component.

Local context (S_l): Local context prior is defined following the Retinex model (*i.e.* change in chromaticity implies change in reflectance). We use this prior in the logarithmic form [25] and substitute $\log\rho = \log I - \log\sigma$ to obtain:

$$S_l = \sum_i \sum_{j \in \mathcal{N}_{3\times3}} w^l_{ij} \big((\log p_i - \log\sigma_i) - (\log p_j - \log\sigma_j)\big)^2$$

where $w^l_{ij} = \exp\big(-\frac{(1-\langle \overline{p_i}\cdot\overline{p_j}\rangle)^2}{t_c^2}\big)\Big[1 + \exp\big(-\frac{p_i^2+p_j^2}{t_b^2}\big)\Big]$. Here $\overline{p_i}$ is pixel chromaticity computed as normalized RGB vector. The first term in the product awards higher value to similarly colored pixel pairs. The second term gives higher weight to pairs with very low intensity values. This reduces color artifacts by suppressing chromatic noise in the dark regions. t_m, t_c and t_b are fixed deviation parameters for weight estimation. We solve this quadratic optimization problem ($\sigma^* = \text{argmin}_\sigma \Psi$) using gradient descent and set $\rho^* = I - \sigma^*$.

2.3 Reflectance Formulation

Unlike our shading formulation (subsection 2.2) which enforces smoothness using L_2 terms, our reflectance formulation enforces color sparsity using L_1 terms. The backbone of this stage is inspired from image flattening work by [7] which uses Split-Bregman method [22] for optimization. For IID they use flattened image as input and perform a series of steps like self-adaptive clustering, Gaussian mixture modeling, boosted tree classification, CRF labeling and L_2 energy minimization. We show that we can use Split-Bregman iterations for direct IID by using proper context priors and alternating between shading and reflectance formulations. In addition to being a direct approach, our method is also more robust to clustering artifacts (Figure 4). Our reflectance formulation is given as:

$$\pi = \gamma_g R_g + \gamma_m R_m + \gamma_l R_l + \gamma_a R_a \tag{5}$$

Here R_g, R_m, R_l and R_a are global, mid-level, local and image approximation terms respectively and γ_g, γ_m, γ_l and γ_a are the associated weights. We use a similar definition for local and global prior weights (v^l and v^g) and have a fixed deviation parameter (t):

$$v_{ij} = \exp\left(-\frac{(\overline{r_i}-\overline{r_j})^2}{2t^2}\right). \tag{6}$$

Here $\overline{r_i}$ is channel normalized CIELab color value with a suppressed luminance [7]. Note that unlike Bi et al. [7], we re-estimate priors in each iteration which gradually leads to IID directly instead of image flattening.

Local context (R_l): We define local reflectance energy term by enforcing the piecewise local image sparsity like Bi et al. [7]:

$$R_l = \sum_i \sum_{j \in \mathcal{N}_{11\times 11}} v^l_{ij} \|R_i - R_j\|_1 = \|\mathbf{Az}\|_1, \tag{7}$$

where R_i represents the reflectance to be computed at pixel position i. This term enforces sparsity on reflectance values using local color information in the form of weights v_{ij} in a 11×11 neighbourhood. This term can be rewritten in matrix form by linearizing the color channels as a single column (z) and assembling a block matrix A of associated pixel weights.

Mid-level context (R_m): As R_l enforces sparsity based only on color similarity in a small local neighbourhood, for mid-level context we enforce sparsity at object level using our selective search features g. For ease of computation, we reduce the dimensions of g to get $\mathring{g}$ using PCA and redefine the weights as :

$$v^m_{ij} = \exp\left(-\frac{(\overline{r_i}-\overline{r_j})^2}{2t^2}\right)\left(-\frac{(\mathring{g}_i-\mathring{g}_j)^2}{2t^2}\right). \tag{8}$$

This prior enforces reflectance sparsity at object level which leads to colour constancy within an object. This captures object level semantics better compared to the local reflectance sparsity constraints which might lead to over flattening due to ambiguity between edges, textures and noise in an image.

$$R_m = \sum_i \sum_{j \in \mathcal{N}_{11\times 11}} v^m_{ij} \|R_i - R_j\|_1 = \|\mathbf{Bz}\|_1. \tag{9}$$

Global context (R_g): The global reflectance prior encodes reflectance similarity at the scene level which is useful in enforcing colour constancy for various instances and occlusion disconnected parts of an object in the scene.

$$R_g = \sum_{i \in Q} \sum_{j \in Q} v^g_{ij} \|R_i - R_j\|_1 = \|\mathbf{Cz}\|_1. \tag{10}$$

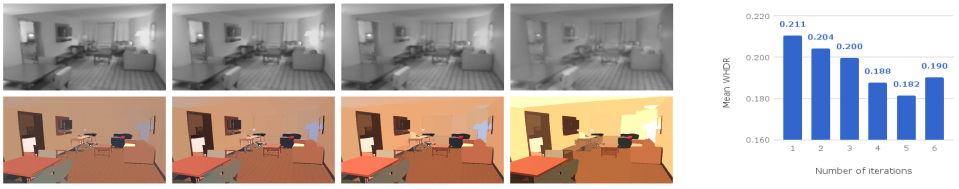


Figure 2: Iteration analysis (From L to R): Shading formulation results (σ^*) and reflectance formulation results (R^*) for iterations $k = 1,3,5,7$. Notice how shading gets 'smoother' while reflectance becomes 'flatter'. The graph shows iterative WHDR reduction for the image with minimum at $k = 5$.

We define Q as the set of representative pixels obtained from each MCG segmentation by ranking all the pixels in a segmentation according to minimum distance from the mean.

Image approximation (R_a): This term enforces continuity between the two stages by forcing the reflectance estimate from the current stage to be similar to the intermediate reflectance solution from the previous shading formulation stage:

$$R_a = \|R_i - \rho\|_2^2 = \|\mathbf{z} - \rho^*\|_2^2 = \|\mathbf{D}\|_2^2. \tag{11}$$

2.4 Iterations and Updates

Using Equation 7, Equation 9, Equation 10 and Equation 11 we can restate Equation 5 in matrix form as:

$$\pi = \|\mathbf{Az}\|_1 + \|\mathbf{Bz}\|_1 + \|\mathbf{Cz}\|_1 + \|\mathbf{D}\|_2^2 \tag{12}$$

This is an $L_1 - L_2$ minimization problem and can be solved using Split-Bregman iterations [22] by introducing intermediate variables $\mathbf{b}$ and $\mathbf{d}$ which reformulates the equation as:

$$\mathbf{z} = \underset{\mathbf{z}}{\operatorname{argmin}} \left(\|\mathbf{D}\|_2^2 + \theta \left(\|\mathbf{d_1} - \mathbf{Az} - \mathbf{b_1}\|_2^2 + \|\mathbf{d_2} - \mathbf{Bz} - \mathbf{b_2}\|_2^2 + \|\mathbf{d_3} - \mathbf{Cz} - \mathbf{b_3}\|_2^2 \right) \right) \tag{13}$$

Here θ balances the contribution from reflectance sparsity priors vs. prior for shading consistency from previous stage. We recompute priors after each iteration for the two formulations based on the current values of σ^* and ρ^* and gradually update the contribution of various weighing parameters (λ, γ and θ), increasing the effect of mid-level and global priors and reducing the effect of local priors over the course of iterations. We estimate the value of all the parameters empirically by tuning for optimal results over a small subset of images.

3 Analysis

Framework analysis: In Figure 2 we show quantitative and qualitative performance of our method for a sample image over successive iterations. Notice how as per the intended design of our framework the reflectance component from our second formulation gradually gets more 'flattened' while shading from the first formulation becomes smoother . Split-Bregman method uses reconstruction error as the stopping criterion [7, 22] but in our case it cannot be directly used to quantify IID performance. Hence we empirically estimate the value of k. Considering various scene and lighting settings we observed that overall our algorithm achieves peak perceptual and quantitative performance for $k = 5$ which can be seen in the WHDR vs. iterations graph in Figure 2. Still better performance could be obtained if IID

Figure 3: Qualitative results on IIW (in each set from L to R): Original image, reflectance and shading. Notice separation of shadows and highlights in shading and color sparsity in reflectance component.

Ablation Analysis

Variant	Shading priors	Reflectance priors	Mean WHDR
v1	S_l	$R_l + R_m + R_g$	32.32
v2	$S_l + S_g$	$R_l + R_m + R_g$	21.99
v3	$S_l + S_m$	$R_l + R_m + R_g$	18.15
v4	$S_l + S_g + S_m$	R_l	23.86
v5	$S_l + S_g + S_m$	$R_l + R_g$	23.81
v6	$S_l + S_g + S_m$	$R_l + R_m$	18.21
v7	$S_l + S_g + S_m$	$R_l + R_m + R_g$	18.19

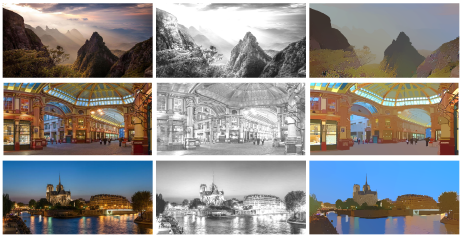


Table 1: Ablation analysis and our results on challenging Internet images highlighting generality of our method for a variety of scene types and light settings.

quality could be approximated for each image separately without ground truth information. But devising such a metric is non-trivial and beyond the scope of this paper. From our experiments we observed that manually selecting optimum k for each image significantly reduces the error.

Ablation study: In order to highlight the significance of various context priors, we conducted ablation study (Table 1) using different variants of our framework formed by combining different prior terms. The study was conducted on a small set of randomly selected 289 images from IIW dataset. Variant v1 is essentially iterative Retinex model based smoothing followed by image flattening. Similarly v4 is only local L_1 flattening performed on top of L_2 shading formulation. Addition of other context priors on top of these basic variants successively improves the performance proving the significance of these priors. In v2 and v5, we introduce the global context priors, leading to significant improvement in performance. Notice the large error drop from v1 $\rightarrow$ v2 is due to our global semantic priors based on RCNN features (f_b) computed on a fixed grid. In v3 and v6 we introduce mid-level context priors using selective search features (g_i) computed using flexible regions, which further leads to significant error reduction from v2 $\rightarrow$ v7 and v5 $\rightarrow$ v7. This shows the utility of our semantic priors at various context levels. Overall the combination of all these priors gives the best IID results both qualitatively and quantitatively.

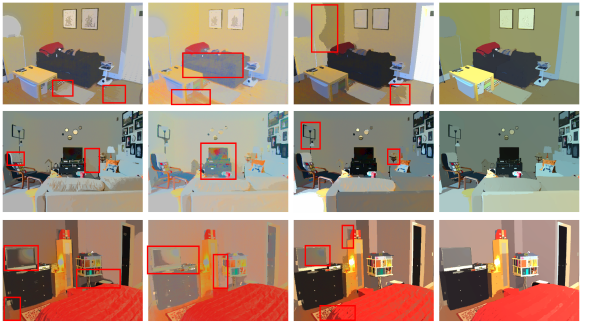
4 Results

All our results are generated using a 5^{th} generation Intel i7 3.30 GHz desktop processor. Most of our prototype implementation is in Matlab with a few sections in C++ suggesting a significant scope of improvement for runtime efficiency. We show the results of our method on the IIW dataset in Figure 3. Notice separation of shadows and illumination from light sources to the shading component and the color consistency in the reflectance component. To explore the generality of our method beyond IIW dataset (only indoor scenes), we also experimented with diverse images from the Internet which are shown alongside Table 1. Our method can work in varying scene types with high complexity and diverse lighting.

We compare our method with other contemporary IID methods which encode scene information in terms of IID priors ([7], [51] and [6]). The results are shown in Figure 4 for the entire IIW dataset (red) and the test-split used in [39] (blue). As [51] uses most of IIW dataset for training, we show their results only on the test-split. The scores are reported as mentioned in the respective papers or downloaded from the respective project webpages. We also compare our method with three baselines. *Baseline 1* is where only shading formulation is optimized and similarly *Baseline 2* is only with the reflectance optimization. Notice that our *Baseline 2* performs better than both [51] and [6] which highlights the strength of our reflectance priors. Also in order to show how different it is from the underlying image flattening framework, we have *Baseline 3* which is computed directly on the results of edge preserving smoothing from [7]. As can be seen from the graph in Figure 4, our method achieves significant error reduction in comparison to both [6] and [51] on both the test-split and the full dataset (WHDR of **17.72** vs. 20.6 and 19.9 respectively on the [39] test-split). Our method is competitive in performance to both [7] and [41] (with WHDR 17.67 and 17.69 respectively) but with lesser artifacts in reflectance results (Figure 4). Additional comparisons with previous IID methods like [50] and [19], with WHDR as 23.20 and 25.46 respectively, are not shown in graph for the sake of clarity. Also note that in our direct method we do not need to perform separate clustering, classification or CRF labeling steps. Our semantic priors lead to consistent reflectance values with lesser patchy artifacts. Furthermore our results handle chromatic noise much better as can be seen in the reflectance of dark regions.

In parallel to our work in this paper, there are three recent direct deep learning solutions by Li and Snavely [35], Bi et al. [8] and Fan et al. [18] with respective WHDR scores as 20.3, 17.18 and 15.8 (joint training) on the [39] split. In [35] and [8], authors introduce new varying illumination datasets and use the illumination invariant property of reflectance for IID. In [18], authors take inspiration from Nestmeyer and Gehler [41] and perform guidance filtering within the CNN framework rather than a separate post processing step which leads to significant error reduction. Based on this observation, we think that properly incorporating semantic information (perhaps in the form of region proposals or masks) within the deep network architecture would further improve the IID performance. Even with our current framework, if we allow for manual tuning of k parameter for each image, chosen based on image complexity (textures, colours, lighting), the error could be reduced to 15.4. Our observations are in-line with the conclusions provided by Nestmeyer and Gehler [41] that using explicit prior knowledge could significantly improve IID performance and future end-to-end deep learning IID solutions could harness these priors for improved results.

Limitations and Future Work: Also while our priors work on varied scenes and generate lesser artifacts, in few cases it is difficult for us to distinguish sharp shadows and highlights from reflectance variations. Finer textures of similar colour as that of the object, persist



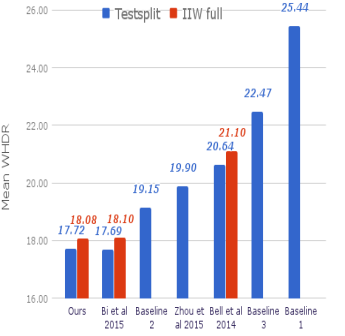


Figure 4: Qualitative comparisons (L to R): Reflectance from [6], [51], [7] and our method. Quantitative resuts: WHDR on testsplit from [39] and the complete dataset [6] (More results in supplementary).

in shading component due to ambiguity in differentiating local illumination changes with such textures (this is not an problem with differently coloured textures). These issues are not unique to our method and are also observed in several other solutions [9]. Still our novel object semantic priors and alternating iterative model design leads to perceptually better decompositions for a large variety of scene and diverse lighting settings. Additional results on IIW, Internet images and qualitative comparisons are included in the supplementary material.

Discounting the training time, deep learning based solutions generally run faster during testing in comparison to energy based optimization methods. Hence the unoptimized prototype implementation of our method is slower compared to other methods (few seconds vs. minutes) but this could be significantly improved with better implementation and parallelization. Additionally, in order to automatically assign the value of total number of iterations k based on the lighting and scene complexity, in future we would like to explore the problem of learning a performance metric for IID respecting both perceptual and quantitative assessment without ground truth information.

5 Conclusion

In this paper we present new priors which encode class agnostic object semantics using selective search and pre-trained region-based Convolutional Neural Network features. We encode these priors by analyzing scene at three hierarchical context levels and use an integrated optimization framework for single image intrinsic image decomposition without requiring any additional optimization steps. Our system has two alternating optimization formulations with competing strategies: first focusing on shading smoothness and the second on reflectance sparsity. We highlight the effectiveness of our strategy and semantic priors with supporting qualitative and quantitative experimentation and results. We hope our work would draw attention of wider research community towards the utility of semantic priors and hierarchical analysis for the problem of intrinsic image decomposition and would lead to a better end-to-end deep learning architecture incorporating these insights.

Acknowledgement: We would like to thank Tata Consultancy Services for supporting Saurabh Saini through Research Scholarship Program (TCS RSP) during the project.

References

[1] P. Arbeláez, J. Pont-Tuset, J. Barron, F. Marques, and J. Malik. Multiscale combinatorial grouping. In *Computer Vision and Pattern Recognition*, 2014.

[2] Jonathan T. Barron. Shape, albedo, and illumination from a single image of an unknown object. In *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, CVPR '12, pages 334–341, 2012.

[3] Jonathan T. Barron and Jitendra Malik. *Color Constancy, Intrinsic Images, and Shape Estimation*, pages 57–70. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012. ISBN 978-3-642-33765-9.

[4] Jonathan T. Barron and Jitendra Malik. Intrinsic scene properties from a single rgb-d image. In *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '13, pages 17–24, 2013.

[5] Jonathan T Barron and Jitendra Malik. Intrinsic scene properties from a single rgb-d image. *TPAMI*, 2015.

[6] Sean Bell, Kavita Bala, and Noah Snavely. Intrinsic images in the wild. *ACM Trans. on Graphics (SIGGRAPH)*, 33(4), 2014.

[7] S. Bi, X. Han, and Y. Yu. An L_1 image transform for edge-preserving smoothing and scene-level intrinsic decomposition. *ACM Transactions on Graphics*, 34(4):78, August 2015.

[8] Sai Bi, Nima Khademi Kalantari, and Ravi Ramamoorthi. Deep Hybrid Real and Synthetic Training for Intrinsic Decomposition. In Wenzel Jakob and Toshiya Hachisuka, editors, *Eurographics Symposium on Rendering*. The Eurographics Association, 2018.

[9] Nicolas Bonneel, Balazs Kovacs, Sylvain Paris, and Kavita Bala. Intrinsic decompositions for image editing. *Computer Graphics Forum (Eurographics State of the Art Reports 2017)*, 36(2), 2017.

[10] Adrien Bousseau, Sylvain Paris, and Frédo Durand. User assisted intrinsic images. *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia 2009)*, 28(5), 2009.

[11] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In A. Fitzgibbon et al. (Eds.), editor, *European Conf. on Computer Vision (ECCV)*, Part IV, LNCS 7577, pages 611–625. Springer-Verlag, October 2012.

[12] Robert Carroll, Ravi Ramamoorthi, and Maneesh Agrawala. Illumination decomposition for material recoloring with consistent interreflections. *ACM Trans. Graph.*, 30(4): 43:1–43:10, July 2011.

[13] Jason Chang, Randi Cabezas, and John W. Fisher. *Bayesian Nonparametric Intrinsic Image Decomposition*, pages 704–719. Springer International Publishing, Cham, 2014.

[14] Q. Chen and V. Koltun. A simple model for intrinsic image decomposition with depth cues. In *2013 IEEE International Conference on Computer Vision*, pages 241–248, Dec 2013.

[15] Jia Deng, Wei Dong, Richard Socher, Li jia Li, Kai Li, and Li Fei-fei. Imagenet: A large-scale hierarchical image database. In *In CVPR*, 2009.

[16] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *International Conference in Machine Learning (ICML)*, 2014.

[17] Sylvain Duchêne, Clement Riant, Gaurav Chaurasia, Jorge Lopez-Moreno, Pierre-Yves Laffont, Stefan Popov, Adrien Bousseau, and George Drettakis. Multi-view intrinsic images of outdoors scenes with an application to relighting. *ACM Transactions on Graphics*, 2015.

[18] Qingnan Fan, Jiaolong Yang, Gang Hua, Baoquan Chen, and David Wipf. Revisiting deep intrinsic image decompositions. 2018.

[19] Elena Garces, Adolfo Munoz, Jorge Lopez-Moreno, and Diego Gutierrez. Intrinsic images by clustering. *Computer Graphics Forum (Proc. EGSR 2012)*, 31(4), 2012.

[20] Peter Vincent Gehler, Carsten Rother, Martin Kiefel, Lumin Zhang, and Bernhard Schölkopf. Recovering intrinsic images with a global sparsity prior on reflectance. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*, NIPS'11, pages 765–773, 2011.

[21] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Computer Vision and Pattern Recognition*, 2014.

[22] Tom Goldstein and Stanley Osher. The split bregman method for l1-regularized problems. *SIAM J. Img. Sci.*, 2(2):323–343, April 2009.

[23] Roger Grosse, Micah K. Johnson, Edward H. Adelson, and William T. Freeman. Ground-truth dataset and baseline evaluations for intrinsic image algorithms. In *International Conference on Computer Vision*, pages 2335–2342, 2009.

[24] Jan Hosang, Rodrigo Benenson, and Bernt Schiele. How good are detection proposals, really? In *Proceedings of the British Machine Vision Conference*. BMVA Press, 2014.

[25] Junho Jeon, Sunghyun Cho, Xin Tong, and Seungyong Lee. Intrinsic image decomposition using structure-texture separation and surface normals. In *Proc. 13th European Conference on Computer Vision (ECCV 2014)*, 2014.

[26] Aditya Khosla, Tinghui Zhou, Tomasz Malisiewicz, Alexei Efros, and Antonio Torralba. Undoing the damage of dataset bias. In *European Conference on Computer Vision (ECCV)*, Florence, Italy, October 2012.

[27] Seungryong Kim, Kihong Park, Kwanghoon Sohn, and Stephen Lin. *Unified Depth Prediction and Intrinsic Image Decomposition from a Single Image via Joint Convolutional Neural Fields*, pages 143–159. Springer International Publishing, Cham, 2016.

[28] Naejin Kong, Peter V. Gehler, and Michael J. Black. *Intrinsic Video*, pages 360–375. Springer International Publishing, Cham, 2014.

[29] V. Kwatra, Mei Han, and Shengyang Dai. Shadow removal for aerial imagery by information theoretic intrinsic image analysis. In *Computational Photography (ICCP), 2012 IEEE International Conference on*, pages 1–8, April 2012.

[30] Pierre-Yves Laffont, Adrien Bousseau, Sylvain Paris, Frédo Durand, and George Drettakis. Coherent intrinsic images from photo collections. *ACM Trans. Graph.*, 31(6): 202:1–202:11, November 2012.

[31] Pierre-Yves Laffont, Adrien Bousseau, and George Drettakis. Rich intrinsic image decomposition of outdoor scenes from multiple views. *IEEE Transactions on Visualization and Computer Graphics*, 19(2):210 – 224, February 2013. presented at SIGGRAPH 2012 (Talk and Poster sessions).

[32] Edwin H. Land and John J. McCann. Lightness and retinex theory. *J. Opt. Soc. Am.*, 61(1):1–11, Jan 1971.

[33] Anat Levin, Dani Lischinski, and Yair Weiss. A closed form solution to natural image matting. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 1*, CVPR '06, pages 61–68, 2006.

[34] Yu Li and Michael S. Brown. Single image layer separation using relative smoothness. In *CVPR*, 2014.

[35] Zhengqi Li and Noah Snavely. Learning intrinsic image decomposition from watching the world. In *Computer Vision and Pattern Recognition (CVPR)*, 2018.

[36] C. Liu, J. Yuen, and A. Torralba. Sift flow: Dense correspondence across scenes and its applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33 (5):978–994, May 2011. doi: 10.1109/TPAMI.2010.147.

[37] Xiaopei Liu, Liang Wan, Yingge Qu, Tien-Tsin Wong, Stephen Lin, Chi-Sing Leung, and Pheng-Ann Heng. Intrinsic colorization. *ACM Trans. Graph.*, 27(5):152:1–152:9, December 2008.

[38] Jonathan Long, Ning Zhang, and Trevor Darrell. Do convnets learn correspondence? In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, NIPS'14, pages 1601–1609, Cambridge, MA, USA, 2014. MIT Press.

[39] Takuya Narihira, Michael Maire, and Stella X. Yu. Learning lightness from human judgement on relative reflectance. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2965–2973, 2015.

[40] Takuya Narihira, Michael Maire, and Stella X. Yu. Direct intrinsics: Learning albedo-shading decomposition by convolutional regression. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, pages 2992–2992, 2015.

[41] Thomas Nestmeyer and Peter V Gehler. Reflectance adaptive filtering improves intrinsic image estimation. In *Computer Vision and Pattern Recognition (CVPR)*, 2017.

[42] Saurabh Saini, Parikshit Sakurikar, and P J Narayanan. Intrinsic image decomposition using focal stacks. In *Proceedings of the Tenth Indian Conference on Computer Vision, Graphics and Image Processing*, ICVGIP '16, pages 88:1–88:8, New York, NY, USA, 2016. ACM.

[43] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. Cnn features off-the-shelf: An astounding baseline for recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2014.

[44] Evan Shelhamer, Jonathan T. Barron, and Trevor Darrell. Scene intrinsics and depth from a single image. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*, December 2015.

[45] Li Shen, Chuohao Yeo, and Binh-Son Hua. Intrinsic image decomposition using a sparse representation of reflectance. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(12): 2904–2915, December 2013.

[46] Marshall F. Tappen, William T. Freeman, and Edward H. Adelson. Recovering intrinsic images from a single image. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(9):1459–1472, September 2005.

[47] A. Torralba and A. A. Efros. Unbiased look at dataset bias. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '11, pages 1521–1528, 2011.

[48] Yair Weiss. Deriving intrinsic images from image sequences. pages 68–75, 2001.

[49] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3320–3328. Curran Associates, Inc., 2014.

[50] Qi Zhao, Ping Tan, Qiang Dai, Li Shen, Enhua Wu, and Stephen Lin. A closed-form solution to retinex with nonlocal texture constraints. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(7):1437–1444, July 2012. ISSN 0162-8828.

[51] Tinghui Zhou, Philipp Krahenbuhl, and Alexei A. Efros. Learning data-driven reflectance priors for intrinsic image decomposition. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, pages 3469–3477, 2015.

[52] Daniel Zoran, Phillip Isola, Dilip Krishnan, and William T. Freeman. Learning ordinal relationships for mid-level vision. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 388–396, 2015.