Learning Data-driven Reflectance Priors for Intrinsic Image Decomposition

Tinghui Zhou
UC Berkeley

Philipp Krähenbühl
UC Berkeley

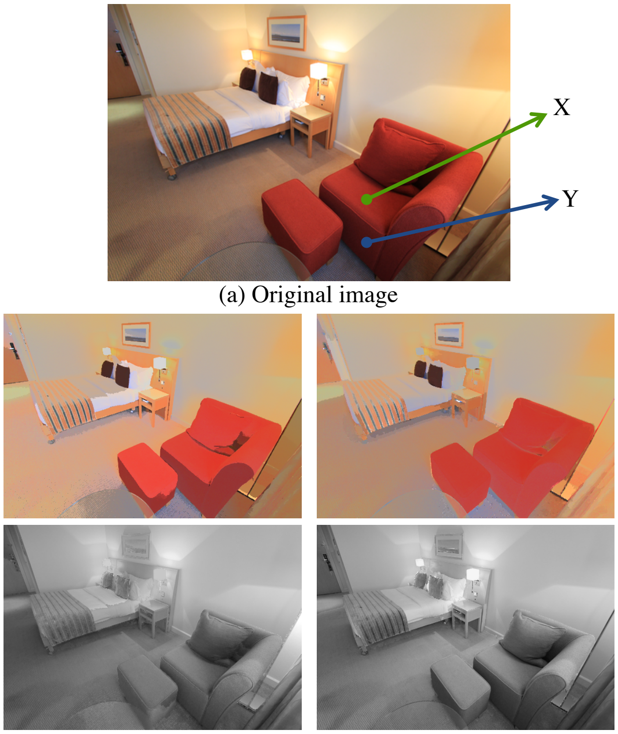
Alexei A. Efros
UC Berkeley

Abstract

We propose a data-driven approach for intrinsic image decomposition, which is the process of inferring the confounding factors of reflectance and shading in an image. We pose this as a two-stage learning problem. First, we train a model to predict relative reflectance ordering between image patches ('brighter', 'darker', 'same') from large-scale human annotations, producing a data-driven reflectance prior. Second, we show how to naturally integrate this learned prior into existing energy minimization frameworks for intrinsic image decomposition. We compare our method to the state-of-the-art approach of Bell et al. *[7] on both decomposition and image relighting tasks, demonstrating the benefits of the simple relative reflectance prior, especially for scenes under challenging lighting conditions.*

1. Introduction

The human visual system is remarkable in its ability to decompose the jumbled mess of confounds that is our visual world into simpler underlying factors. Nowhere is this more apparent than in our impressive ability, even from a single still image, to tease apart the effects of surface reflectance vs. scene illumination. Consider the mini-sofa in Figure 1(a): on one hand, we can see that its seat (point X) is much brighter than its frontal face (point Y), but *at the same time*, we can also clearly tell that they are both "made of the same stuff" and have the same surface reflectance. This is remarkable because, by the time the light has bounced off the sofa toward the eye (or the camera), the contributions of reflectance and illumination have been hopelessly entangled, which the brain then needs to undo.



(a) Original image

(b) Decomposition by Bell *et al*.

(c) Our decomposition

Figure 1: Given an image (a), people have no trouble disentangling the confounding factors of reflectance and shading: we can see that X is much brighter than Y, but *at the same time*, we can also clearly tell that they are both "made of the same stuff" and have the same surface reflectance. Our algorithm (c) automatically decomposes (a) into a reflectance image (c,top) and a shading image (c,bottom). Note how the mini-sofa is a uniform red in our reflectance image, compared to (b) state-of-the-art algorithm of Bell *et al*. [7].

In computer vision, the decomposition of an image into reflectance (albedo) and illumination (shading) maps is usually, if somewhat inaccurately, referred to as the *intrinsic image decomposition* [6][1]. The intrinsic image model states that the observed luminance image is the product of the reflectance image times the shading image. Clearly, the inverse problem of inferring the underlying reflectance and shading images is ill-posed and under-constrained in this pure form since any given pixel intensity could be explained equally well by reflectance or shading [2]. To address this,

[1] The original formulation of Barrow and Tenenbaum [6] also includes other factors, such as depth, orientation, occlusion, transparency, etc

additional constraints (priors) are typically imposed on the decomposition process to capture the statistical and/or physical regularities in natural images. However, those priors are typically hand-crafted and overly weak. For example, one popular prior proposed originally in the Retinex algorithm of Land and McCann [18] assumes that large intensity gradients correspond to reflectance edges, while low-frequency changes are mainly due to shading. While this prior works well in many cases, it fails in the presence of strong shadows, sharp changes in surface orientation, and smoothly-varying planar textures. Since then, many other clever priors have been proposed, including texture statistics [20, 23], shape, albedo, and illumination [3–5], meso- and macro-scales of shading [19], chromaticity segmentation [10], sparsity on reflectances [11, 24], etc., or combination thereof [7], in the hopes of finding the silver bullet which could *fully explain* the intrinsic image phenomenon, but to date none has emerged. One is faced with the possibility that there might not exist a simple, analytic prior and that a more data-driven approach is warranted.

In this paper we propose to learn priors for intrinsic image decomposition directly from data. Compared to other work that trains a reflectance vs. shading classifier on image patches (e.g. [26, 27]), our main contribution is to train a *relative* reflectance prior on *pairs* of patches. Intuitively, the goal is to learn to detect surface regions with similar reflectance, even when their intensities are different. We take advantage of the recently released Intrinsic Images in the Wild (IIW) database of Bell *et al*. [7], in which a large set of relative reflectance judgments are collected from human subjects for a variety of real-world scenes. Other contemporary work, developed independently, have also employed the IIW dataset. Narihira *et al*. [22] use the IIW dataset to learn a perceptual lightness model. The key difference is that we not only learn a relative reflectance prior from pairwise annotations, but also utilize it for intrinsic image decomposition. In these same proceedings, Zoran *et al*. [31] use a similar approach to ours to estimate ordinal relationships between pairs of points, but globalizes them with a different energy optimization.

Our relative reflectance model is an end-to-end trained convolutional neural network that predicts a probability distribution over the relative reflectance (‘brighter’, ‘darker’, ‘same’) between two query pixels. We show how to naturally integrate this learned prior into existing energy minimization frameworks for intrinsic image decomposition, and demonstrate the benefits of such relative reflectance priors, especially for scenes under challenging illumination conditions.

2. Learning a model of reflectance

Let $r_i \in \mathcal{R}$ be a reflectance estimate at pixel i, where $\mathcal{R}$ is the set of all reflectance values in a scene. For two reflectance values $r_i, r_j \in \mathcal{R}$ let $r_i < r_j$ denote that reflectance r_i is darker than reflectance r_j, and $r_i = r_j$ means that the reflectances are roughly equivalent.

Estimating reflectance directly is hard and usually requires a specialized sensor, such as a photometer. Not even the human visual system can infer absolute reflectance reliably (see Adelson [1] for examples). Humans are much better at estimating relative reflectance between two point r_i and r_j in a scene [7]. We follow this intuition and learn a classifier that predicts this relative reflectance between different parts of a scene in Section 2.1. However, just like human reflectance estimates, this classifier might not be globally consistent. Section 2.2 recovers the globally consistent reflectance estimate following our relative estimates. We then use this global reflectance model in Section 3 to guide an intrinsic image decomposition.

2.1. Relative reflectance classifier

For two pixels i and j in a scene, our goal is to estimate the relative reflectance between them as being equal $r_i = r_j$, darker $r_i < r_j$ or brighter $r_i > r_j$. Our relative reflectance classifier is a multi-stream convolutional neural network (see Fig. 2), accounting for 1) local features around pixel i, 2) local features around pixel j, 3) global scene features of the input image, and 4) spatial coordinates of both input pixels, respectively. The network weights are shared between the two local feature extraction streams. All features are then concatenated, and fed through three fully-connected layers that predict classification scores over the relative reflectance labels (‘same’, ‘darker’, ‘brighter’). Each convolution and fully-connected layer (except for the last prediction layer) is followed by a rectified linear unit.

We train this network from scratch using the pairwise human judgments of the Intrinsic Images in the Wild dataset [7] and millions more obtained through symmetry and transitivity properties of the original annotations (see Section 4.1 for details on data augmentation). The network is learned end-to-end in CAFFE [13] using a softmax loss. Our network outperforms all state-of-the-art methods in terms of relative reflectance predictions, as we will show in Section 4. However the resulting predictions are not always globally consistent. This is in part due to inconsistencies in the human-annotated training data. Roughly 7.5% of all training annotations are inconsistently labeled [7] and our network learns part of that inconsistency.

Next, we show how to recover a globally consistent reflectance estimate from the noisy pairwise predictions produced by the classifier.

2.2. Globally consistent reflectance estimate

The network output gives an estimate for the relative reflectance between a pair of pixels i and j. Let $w_{=,i,j}$, $w_{<,i,j}$ and $w_{>,i,j}$ be the classifier score of ‘same’, ‘darker’, and

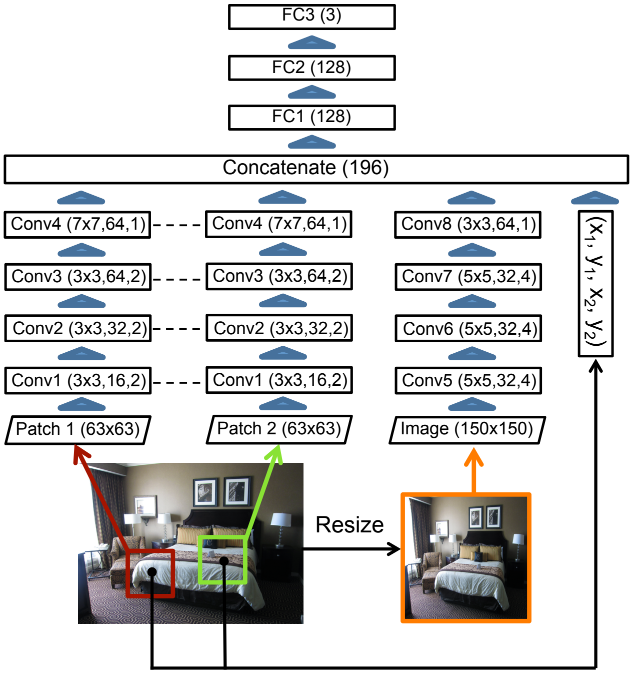


Figure 2: Our multi-stream network architecture for relative reflectance prediction. The network weights are shared between the local feature extraction streams. Features extracted from all four streams are fed through three fully-connected layers for final relative reflectance prediction (see Section 2.1 for more details).

'brighter', respectively. A higher score enforces a larger consistency for a specific pairwise comparison. We constrain all weights to be non-negative, and formulate global reflectance estimation as a constrained optimization problem, where each classifier output imposes a pairwise constraint on the global ordering

$$\begin{aligned} \underset{\boldsymbol{r},\varepsilon}{\text{minimize}} \quad & \sum_{i,j\in\mathcal{E}} \sum_{o\in\{=,<,>\}} w_{o,j,i}\xi_{o,i,j} \\ \text{subject to} \quad & r_i \le r_j + \xi_{=,i,j} \\ & r_j \le r_i + \xi_{=,i,j}, \\ & r_i \le r_j + \xi_{<,i,j}, \\ & r_j \le r_i + \xi_{>,i,j}, \\ & \xi \ge 0. \end{aligned} \tag{1}$$

Here, ξ is a slack variable that tries to enforce all constraints as well as possible. All pairwise reflectance measures are evaluated on a set of sparse edges $\mathcal{E}$.

This constrained optimization naturally translates into a global energy minimization:

$$E(\boldsymbol{x}) = \sum_{i,j\in\mathcal{E}} \sum_{o\in\{=,<,>\}} w_{o,j,i}\, \mu_o(r_i, r_j), \tag{2}$$

where $\mu_<$, $\mu_>$, and $\mu_=$ penalizes the disagreement between our classifier and the globally consistent ranking.

For objective Eq. 1 this translates into a hinge loss, that penalizes the degree to which the consistent reflectance estimate disagrees with our classifier:

$$\begin{aligned} \mu_=(r_i, r_j) &= \xi_{=,i,j} = |r_i - r_j| \\ \mu_<(r_i, r_j) &= \xi_{<,i,j} = \max(r_i - r_j, 0) \\ \mu_>(r_i, r_j) &= \xi_{>,i,j} = \max(r_j - r_i, 0). \end{aligned}$$

For continuous values r_i the energy minimization 2 is convex. For discrete values r_i it can be expressed as a binary submodular problem on an extended sparsely connected graph [16]. We use GraphCuts to globally optimize it [9].

While objective 1 computes a global ordering on the reflectance, it does not provide information about the absolute reflectance in an image. In the next section we will show how to incorporate the reflectance prior into a standard intrinsic image decomposition pipeline to recover an absolute estimate of reflectance.

3. Intrinsic image decomposition

We start out with the intrinsic image decomposition framework of Bell *et al.* [7]. Given an input image I, their system recovers a reflectance image $\boldsymbol{r}$ and shading image $\boldsymbol{s}$. They model intrinsic image decomposition as an energy minimization in a fully connected CRF [15].

$$E(\boldsymbol{s}, \boldsymbol{r}) = \sum_i \psi_i(r_i, s_i) + \sum_{i>j} \psi^r_{ij}(r_i, r_j) + \psi^s_{ij}(r_i, r_j), \tag{3}$$

where ψ_i is a unary term that captures some lightweight unary priors on absolute shading intensity or chromaticity of the reflectance as an L1 norm between the original image and the estimated properties. The unary term also constrains the reflectance and shading to reconstruct the original image. Most of the heavy lifting of the model is done by the pairwise terms ψ^r and ψ^s that enforce smoothness of reflectance and lighting respectively.

The pairwise shading term is modeled as a fully connected smoothness prior:

$$\psi^s_{ij}(r_i, r_j) = (s_i - s_j)^2 \exp\left(-\beta_1 (\boldsymbol{p}_i - \boldsymbol{p}_j)^2\right),$$

where $\boldsymbol{p}_i$ is the position of a pixel i, and β_1 is a parameter controlling the spatial extent of the prior. This prior captures the intuition that the shading varies smoothly over smooth surfaces.

The pairwise reflectance term is modeled as a color sensitive regularizer encouraging pixels with a similar color value in the original image to take a similar reflectance:

$$\psi^r_{ij}(r_i, r_j) = |r_i - r_j| \exp\left(-\beta_2 (\boldsymbol{p}_i - \boldsymbol{p}_j)^2 - \beta_3 (\boldsymbol{I}_i - \boldsymbol{I}_j)^2\right),$$

where $\boldsymbol{I}_i$ is color value of a pixel i, and β_2 and β_3 control the spatial and color extent of the prior. This reflectance term is quite arbitrary, as original color values are usually not a good sole predictor of reflectance. In the rest of this section we will show how to replace this term with our data-driven pairwise reflectance prior.

The overall energy $E(\boldsymbol{s}, \boldsymbol{r})$ is optimized using an alternating optimization for $\boldsymbol{s}$ and $\boldsymbol{r}$. The reflectance term $\boldsymbol{r}$ is optimized using the mean-field inference algorithm of Krähenbühl and Koltun [15], while the shading term is optimized with iteratively reweighted least squares (IRLS).

3.1. Data-driven reflectance prior

We now show how to incorporate our relative reflectance classifier into the mean-field inference for reflectance. Specifically we define our new pairwise term as

$$\psi^r_{ij}(r_i, r_j) = \sum_{o \in \{=,<,>\}} \mu_{o,i,j}(r_i, r_j) w_{o,i,j}, \tag{4}$$

The main difficulty here is to evaluate the pairwise term densely over the image. The mean-field inference algorithm relies on an efficient evaluation of $\tilde{Q}_i(r_i) = \sum_j \sum_{r_j} \psi^r_{ij}(r_i, r_j) Q_(r_i)$, which is known as message passing. This message passing step naturally decomposes into a matrix multiplication with μ_o and a filtering term with w_o. The matrix multiplication can be evaluated efficiently as it is independent for each pixel and scales linearly in the number of pixels. The filtering step on the other hand requires an exchange of information between each pair of pixels in the image. Krähenbühl and Koltun [15] showed that for a Gaussian pairwise term the filter can be approximated efficiently. The same Gaussian pairwise term is used in the original model of Bell *et al.* [7]. In our model this filter is no longer a simple Gaussian kernel, but guided by the output of a classifier. The filtering has the following form

$$\hat{Q}^{(o)}_i(l) = \sum_j w_{o,i,j} Q_j(l), \tag{5}$$

for each comparison $o \in \{<, >, =\}$. For our data-driven pairwise term we would need to evaluate a classifier densely over each pair of pixels in the image, which is computationally intractable for more than a few thousand pixels.

However, the classifier output is quite low rank. If we denote $|\mathcal{R}|$ as the number of unique reflectance values in a scene, which is usually small [11, 24], then the output of an ideal classifier is of at most rank $|\mathcal{R}|$. This comes from the fact that each reflectance value $r \in \mathcal{R}$ forms a binary basis B, with a value of $B_{i,r} = 1$ if pixel i takes reflectance r, and $B_{i,r} = 0$ otherwise. Thus any ideal classifier output can be expressed as a product of $B\tilde{W}_o B^\top$, where $\tilde{W}_o$ is a $|\mathcal{R}| \times |\mathcal{R}|$ matrix describing the weighting between different reflectance values. Any rank beyond this can be attributed to noise or inconsistencies in the classifier. We measured the rank of the classifier matrix by randomly sampling $K = 500$ points in the image and computing the full pairwise term between those points. This results in a $K \times K$ pairwise comparison matrix. We never encountered this classifier matrix to be of rank more than 100. This suggests that the low rank approximation models $w_<$, $w_=$ and $w_>$ well.

3.2. Nyström approximation

We use Nyström's method [17] to approximate w_o. The main caveat with Nyström is that it requires a symmetric pairwise comparison matrix w_o. While the equality constraint matrix $w_=$ is symmetric, the inequality matrices are not $w_> = w_<^\top$. We address this by rearranging all classifier outputs in a larger comparison matrix W:

$$W = \begin{pmatrix} w_{=,1,1} & w_{>,1,1} & w_{=,1,2} & w_{>,1,2} & \cdots \\ w_{<,1,1} & w_{=,1,1} & w_{<,1,2} & w_{=,1,2} & \cdots \\ w_{=,2,1} & w_{>,2,1} & w_{=,2,2} & w_{>,2,2} & \cdots \\ w_{<,2,1} & w_{=,2,1} & w_{<,2,2} & w_{=,2,2} & \cdots \\ \cdots & \cdots & \cdots & \cdots & \end{pmatrix}$$

This extended matrix is symmetric and can be well approximated using Nyström's method. It is still low rank, as the three submatrices it comprises of are all low rank. We can compute the filtering in Eq. 5 by multiplying W with a vector $[Q_1(l), 0, Q_2(l), 0, Q_3(l), \ldots]^\top$ and extracting every other elements from it.

The Nyström approximation samples $2K$ rows from matrix W. Let C denote those sampled rows. We always sample pairs of consecutive rows, to not introduce a bias towards any of the operations $=, <$ or $>$, Nyström then approximates the dense pairwise classifier matrix as

$$W \approx C D^+ C^\top,$$

where D is a $K \times K$ matrix corresponding to the dense pairwise classifier scores between all sampled points, and $^+$ refers to the pseudo-inverse. We sample $K = 64$ on a regular grid, which allows us to compute the matrices C and D within 10 seconds including the classifier evaluation. The Nyström approximation allows us to compute a message passing step within a few hundred milliseconds, while a naive evaluation would take multiple days to compute.

In summary, we evaluate the pairwise reflectance classifier from K sampled points to all other points in the image. The Nyström approximation then allows us to approximate a fully-connected dense pairwise comparison matrix using those few samples, which in turn allows for a natural integration into the fully connected CRF framework of Krähenbühl and Koltun. Notice that Nyström approximation for dense CRF has recently been explored in [29]. However, [29] merely approximates the commonly used Gaussian kernel, while we show how to integrate a more general output of a classifier into the dense CRF framework.

4. Experiments

In this section, we evaluate the performance of each component of our pipeline using two data sources: 1) Intrinsic Images in the Wild (IIW) dataset [7] and 2) Image Lighting Composition (ILC) dataset [8]. Our main baseline is the state-of-the-art intrinsic image decomposition algorithm by Bell *et al.* [7]. All models are trained and evaluated on the dataset split of Narihira *et al.* [22].

4.1. Data augmentation

IIW dataset provides $875,833$ comparisons across $5,230$ photos, which we extensively augment by exploiting the symmetry and transitivity of the comparisons. The augmentation not only helps reduce overfitting (as shown in Section 4.2), but also generates pixel pairs that are spatially distant from each other (in contrast to ones originally derived from edges of a Delauney triangulation [7]). We create the augmented training and test annotations as follows:

1. Remove low-quality comparisons with human confidence score < 0.5.
2. For each remaining pairwise comparison (r_i, r_j), augment the annotation for (r_j, r_i) by either flipping (if $r_i \neq r_j$) or keeping (if $r_i = r_j$) the sign.
3. For any unannotated pair of reflectances (r_i, r_j) that share a comparison with r_k, we augment it using the following rules: 1) $r_i = r_j$, iff $r_i = r_k$ and $r_j = r_k$ for all connected r_k; 2) $r_i > r_j$, iff $r_i \geq r_k > r_j$ or $r_i > r_k \geq r_j$; 3) $r_i < r_j$, iff $r_i < r_k \leq r_j$ or $r_i \leq r_k < r_j$. If any pairwise comparisons are inconsistent we do not complete them. This step is done repetitively for each image until no further augmentation is possible.

Our augmentation generates $22,903,366$ comparisons in total, out of which $18,621,626$ are used for training and $4,281,740$ for testing.

4.2. Network performance

We use ADAM [14] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, an initial learning rate of 0.001, step size of $20,000$, a step multiplier $\gamma = 0.8$. We train with mini-batches of 128 pairs and weight decay of 0.002.

For evaluation, we first use the same *weighted human disagreement rate* (WHDR) metric as [7] on the test split. WHDR measures the percent of human judgments that a model incorrectly predicts, weighted by the confidence of each judgment. Note that the human judgments are not necessarily consistent in the IIW dataset as human performance using this metric is 7.5 [7]. As shown in Table 2, our full model trained on the augmented data performs the best with WHDR $= 15.7$.

Data source Metric	Original WHDR	Augmented Error Rate
Bell *et al.* [7]	20.6	27.9
Retinex-Color [12]	26.9	29.3
Retinex-Gray [12]	26.8	30.5
Garces *et al.* [10]	24.8	29.9
Shen and Yeo [25]	32.5	34.2
Zhao *et al.* [30]	23.8	31.1
Narihira *et al.* [22]	18.1	36.6
Local	16.6	25.8
Local + Spatial	16.1	25.1
Local + Spatial + Global	**15.7**	**24.6**
Local + Spatial + Global (Orig.)	17.3	32.4

Table 1: Performance on the IIW dataset [7] measured by WHDR (left) on the original, locally-connected comparisons and Error rate (right) on our augmented, potentially long-range comparisons. The bottom four rows correspond to our models trained with different components: local features only, local and spatial features, full network, and full network trained on original IIW annotations only.

Additionally, we evaluate the error rate of different algorithms on our augmented annotations. Our full model again obtains the lowest error rate of 24.6. More surprisingly, on this metric other baselines surpass the recent top performer [22]. This is likely due to a subtle bias in the original IIW annotations – spatially close pixels often have the same reflectance. This bias is no longer present in our augmented annotations as they contain more long-range pairs. This is further verified by the performance of our full model trained only on the original annotations: it too does poorly on the augmented data.

Globally consistent reflectance estimate We measure the performance of recovering a globally consistent reflectance estimate with the energy optimization presented in Section 2.2. Specifically, for each test image in the IIW dataset, we build a sparse graph over the annotated pixel pairs, and apply the relative reflectance network to each of the sampled pixels. The predicted scores are then jointly optimized by Eq. 2 using GraphCuts [9] to recover the globally consistent ordering. The recovery performance is measured using WHDR, and we obtain 18.0 over the entire test split. Compared to the direct network output (WHDR $= 15.7$), global ordering recovery loses 2.3 percent of the performance due to the inconsistency and noise of the network output.

Nyström approximation We experimented with different point sampling strategies (including random sampling, spatial grid sampling and Poisson disk sampling) as well as

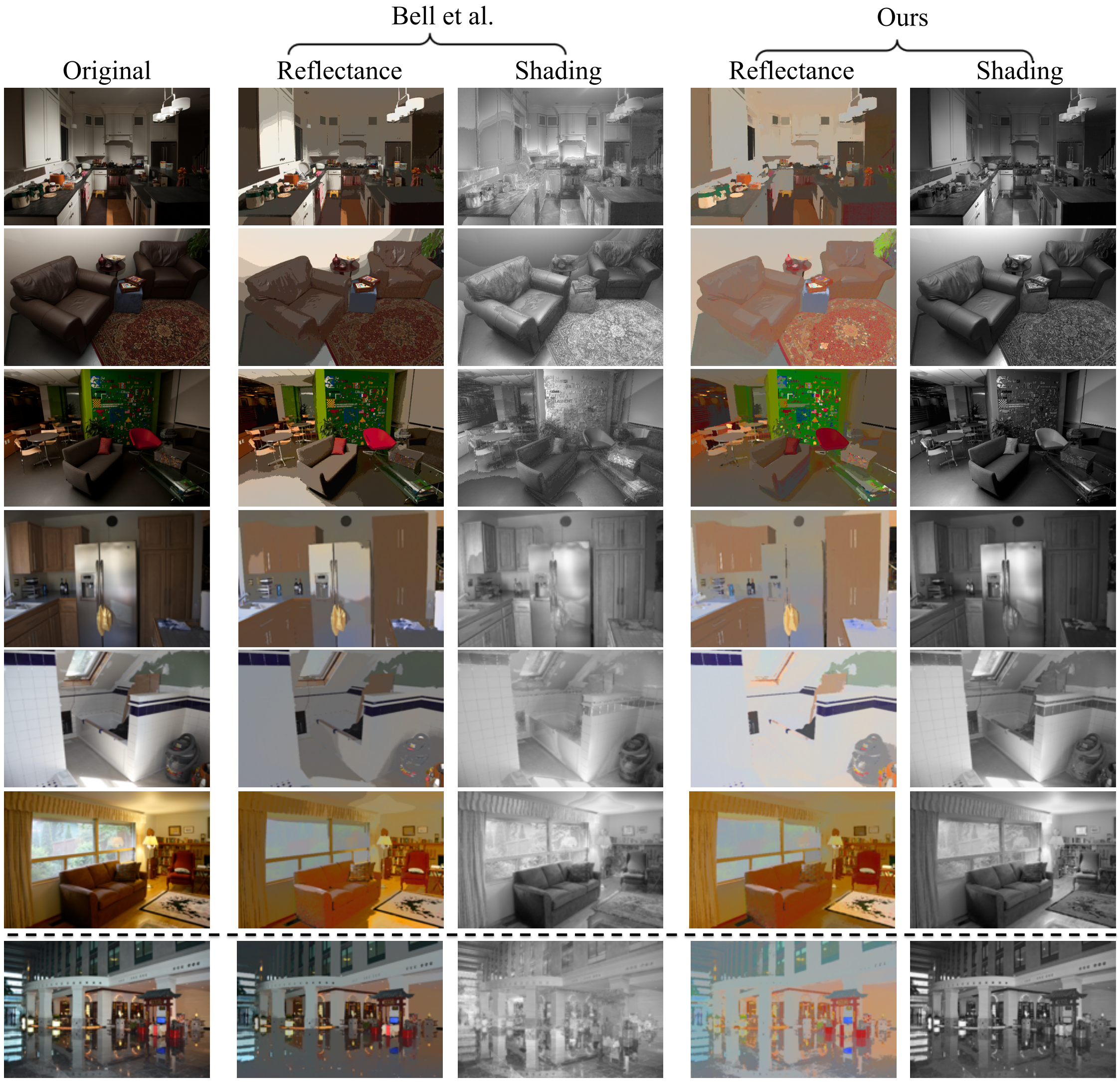


Figure 3: Comparison of intrinsic image decomposition between Bell *et al*. [7] and ours (chrom. + our prior + shading). Rows 1–3 are examples from the ILC dataset [8], and the rest are ones from the IIW dataset [7]. In general, our decomposition tends to distinguish between reflectance and shading boundaries better compared to the baseline, especially under challenging lighting conditions (e.g. Rows 1–3). The last row shows an example where Bell *et al*. outperforms ours due to stronger reflectance smoothness constraints.

different sample sizes, and found that grid sampling with 64 samples to work well. More samples tend to yield better approximation at the cost of computation. The overall WHDR on the IIW test split using Nyström approximated pairwise comparison is 17.2, which is slightly worse than the direct network output (15.7).

4.3. Intrinsic image decomposition

To understand the effect of our reflectance prior on intrinsic image decomposition, we perform an ablation study on several variants of the decomposition framework:

Chromaticity only each pixel being assigned to the reflectance label that is most similar in chromaticity. This

Data source Metric	Original WHDR	Augmented Error Rate
Bell *et al*. [7]	20.6	27.9
Chromaticity only	33.6	38.5
Chrom. + Our prior	22.5	29.6
Chrom. + Our prior + Shading	**19.9**	**27.3**

Table 2: Ablation study on different variants of the decomposition framework. All results are on the test set of IIW.

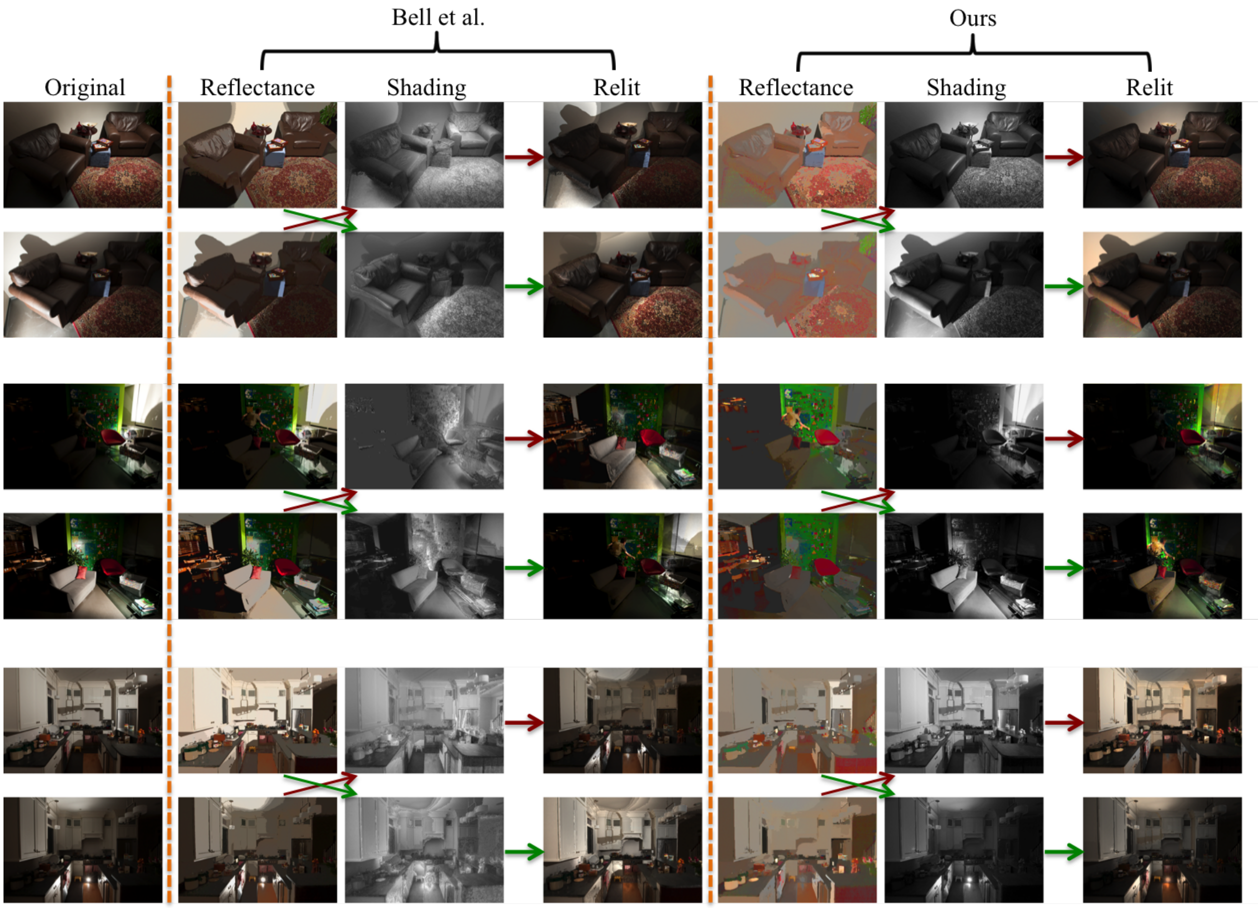


Figure 4: Comparison of relighting results between Bell *et al.* [7] and ours on the variable lighting dataset of [8]. In each row, we construct a relit image from the shading in the same row and the reflectance of the adjacent row. We expect a minimal change in appearance between the original and relit images, since they depict the same scene and thus should share the same reflectance. Our relighting results tend to reconstruct the target images more closely, which also implies better decomposition performance. See Section 4.4 for more details.

simple variant achieves WHDR = 33.6 on the original annotations, and error rate = 38.5 on the augmented data.

Chromaticity + our prior dense CRF with chromaticity similarity as the unary potential and our reflectance prior as the pairwise potential. This variant greatly improves the performance over using chromaticity only with WHDR = 22.5 and error rate = 29.6 on the original and augmented annotations, respectively, indicating the effectiveness of our reflectance prior.

Chromaticity + our prior + shading previous variant with additional shading costs from Bell *et al*. [7]. This variant achieves the best decomposition performance with WHDR = 19.9 and error rate = 27.3. It improves on the decomposition of Bell *et al*. both quantitatively and qualitatively.

We visualize our final decomposition output (chrom. + our prior + shading), and compare with Bell *et al*. [7] in Figure 3 for examples from both IIW dataset and ILC datasets. In general, our decomposition tends to distinguish between reflectance and shading boundaries better than the baseline, especially under challenging lighting conditions (e.g. examples from the ILC dataset). For instance, for the kitchen scene in the first row of Fig. 3, Bell *et al*. failed to separate the shading layer from the reflectance layer correctly, leading to large shadow boundaries (see cupboards and the floor) left over in the reflectance layer. Similarly for the example in row 6 of Fig. 3, Bell *et al*. failed to recognize that the drastic intensity change on the ceiling and floor is due to illumination from the lamp, whereas our decomposition was able to correctly identify the shadows, and attribute them to the shading layer. However, the hand-crafted reflectance smoothness prior still works more favorably in some cases (e.g. the last row of Fig. 3).

	Kitchen	Sofas	Cafe	Mean
Bell *et al.* [7]	8.66	8.39	8.55	8.53
Ours	**6.93**	**6.87**	**6.63**	**6.81**

Table 3: Mean pixel reconstruction error (MPRE) on three illumination varying sequences ($\times 10^{-4}$). Lower is better.

4.4. Robustness to illumination variation

An ideal reflectance model should be invariant to illumination changes. To measure the degree of illumination invariance, we use image sequences of indoor scenes taken by a stationary camera under different lighting conditions provided by [8], and perform relighting experiments on decomposition outputs of our method and Bell *et al.* Specifically, given two images I_A and I_B taken from the same scene and their decomposition $I_A = R_A S_A$ and $I_B = R_B S_B$ respectively, perfect decomposition would imply equal reflectance $R_A = R_B$, and the difference between I_A and I_B is entirely explained by the shading/lighting components S_A and S_B. In other words, for ideal decompositions, we should be able to relight R_A using S_B to reconstruct I_B (and similarly use R_B and S_A to reconstruct I_A). Thus, we propose to use mean pixel reconstruction error (MPRE), $\frac{1}{N^2 P} \sum_A \sum_B \|R_A S_B - I_B\|_2$, for measuring illumination invariance, where N is the number of images, and P is the number of pixels per image. We report the MPRE results for the three indoor scene sequences in Table 3, and a qualitative comparison in Fig. 4. We significantly outperform Bell *et al.* both quantitatively and perceptually.

4.5. Feature visualization

Finally, we visualize the features learned by our relative reflectance network using the t-SNE algorithm [28]. Specifically, we randomly extract $50,000$ patches from the test set of IIW and find a 2-dimensional embedding of their 64-dimensional Conv4 features. Fig. 5 shows this embedding. The overall layout appears to be highly predictive of reflectance (light to dark from top-left to bottom-right). Moreover, it seems to discover some surface or material properties beyond reflectance (see Fig. 5 for more details).

Discussion

One limitation of our paper is that although the learned reflectance prior accounts for most of the decomposition performance, hand-crafted unaries on chromaticity and shading are still used for achieving state-of-the-art results. However, while it is beyond the scope of this paper, we believe hand-crafted unaries can be replaced by learned unaries (c.f. concurrent work of Narihira *et al.* [21]).

Acknowledgements This research is supported by ONR MURI N000141010934, Intel research grant, and Nvidia hardware donation.

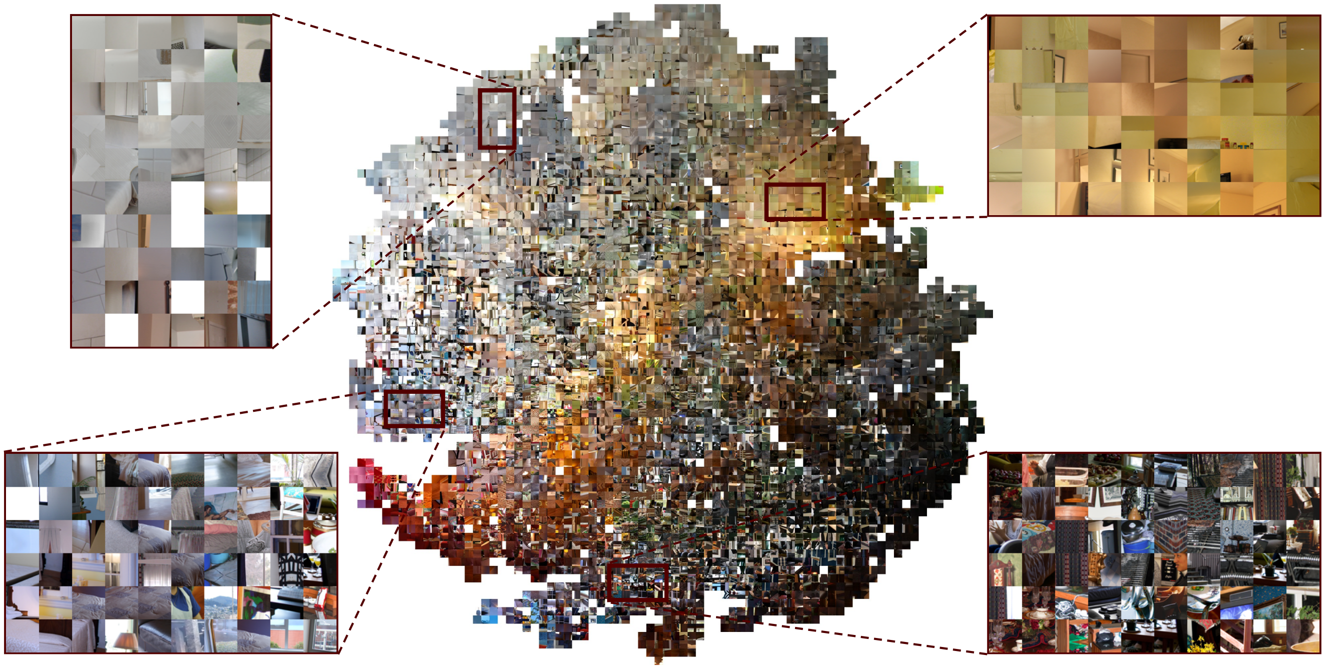


Figure 5: Feature embedding visualized by t-SNE [28]. The learned features are usually highly predictive of surface color (bottom right). More interestingly, our network is also able to coherently group patches based on properties beyond reflectance. For example, the network groups bathroom tiles (top left), wall paper (top right), or cloth surfaces (bottom left), based on material properties or local appearance.

References

[1] E. H. Adelson. Lightness perception and lightness illusions. *The new cognitive neurosciences*, page 339, 2000. 2

[2] E. H. Adelson and A. P. Pentland. The perception of shading and reflectance. *Perception as Bayesian inference*, pages 409–423, 1996. 1

[3] J. Barron and J. Malik. Shape, illumination, and reflectance from shading. *PAMI*, 2015. 2

[4] J. T. Barron and J. Malik. Color constancy, intrinsic images, and shape estimation. In *ECCV*, pages 57–70. 2012. 2

[5] J. T. Barron and J. Malik. Shape, albedo, and illumination from a single image of an unknown object. In *CVPR*, pages 334–341, 2012. 2

[6] H. G. Barrow and J. M. Tenenbaum. Recovering intrinsic scene characteristics from images. *Computer Vision Systems*, pages 3–26, 1978. 1

[7] S. Bell, K. Bala, and N. Snavely. Intrinsic images in the wild. *ACM Transactions on Graphics*, 33(4):159, 2014. 1, 2, 3, 4, 5, 6, 7, 8

[8] I. Boyadzhiev, S. Paris, and K. Bala. User-assisted image compositing for photographic lighting. *ACM Transactions on Graphics*, 32(4):36, 2013. 5, 6, 7, 8

[9] Y. Boykov and V. Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *PAMI*, 26(9):1124–1137, 2004. 3, 5

[10] E. Garces, A. Munoz, J. Lopez-Moreno, and D. Gutierrez. Intrinsic images by clustering. In *Computer Graphics Forum*, volume 31, pages 1415–1424, 2012. 2, 5

[11] P. V. Gehler, C. Rother, M. Kiefel, L. Zhang, and B. Schölkopf. Recovering intrinsic images with a global sparsity prior on reflectance. In *NIPS*, pages 765–773, 2011. 2, 4

[12] 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. 5

[13] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. B. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM Multimedia*, pages 675–678, 2014. 2

[14] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 5

[15] P. Krähenbühl and V. Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *NIPS*, pages 109–117, 2011. 3, 4

[16] M. P. Kumar, O. Veksler, and P. H. S. Torr. Improved moves for truncated convex models. *JMLR*, 12:31–67, 2011. 3

[17] S. Kumar, M. Mohri, and A. Talwalkar. Sampling methods for the nyström method. *JMLR*, 13(1):981–1006, 2012. 4

[18] E. H. Land and J. McCann. Lightness and retinex theory. *JOSA*, 61(1):1–11, 1971. 2

[19] Z. Liao, J. Rock, Y. Wang, and D. Forsyth. Non-parametric filtering for geometric detail extraction and material representation. In *CVPR*, pages 963–970. IEEE, 2013. 2

[20] X. Liu, L. Jiang, T.-T. Wong, and C.-W. Fu. Statistical invariance for texture synthesis. *Visualization and Computer Graphics*, 18(11):1836–1848, 2012. 2

[21] T. Narihira, M. Maire, and S. X. Yu. Direct intrinsics: Learning albedo-shading decomposition by convolutional regression. In *ICCV*, 2015. 8

[22] T. Narihira, M. Maire, and S. X. Yu. Learning lightness from human judgement on relative reflectance. In *CVPR*, 2015. 2, 5

[23] B. M. Oh, M. Chen, J. Dorsey, and F. Durand. Image-based modeling and photo editing. In *ACM Computer graphics and interactive techniques*, pages 433–442, 2001. 2

[24] I. Omer and M. Werman. Color lines: Image specific color representation. In *CVPR*, pages 946–953, 2004. 2, 4

[25] L. Shen and C. Yeo. Intrinsic images decomposition using a local and global sparse representation of reflectance. In *CVPR*, pages 697–704, 2011. 5

[26] M. F. Tappen, E. H. Adelson, and W. T. Freeman. Estimating intrinsic component images using non-linear regression. In *PAMI*, volume 2, pages 1992–1999, 2006. 2

[27] M. F. Tappen, W. T. Freeman, and E. H. Adelson. Recovering intrinsic images from a single image. *PAMI*, 27(9):1459–1472, 2005. 2

[28] L. van der Maaten and G. Hinton. Visualizing high-dimensional data using t-sne. *JMLR*, 2008. 8

[29] P. Wang, C. Shen, and A. van den Hengel. Efficient sdp inference for fully-connected crfs based on low-rank decomposition. In *CVPR*, 2015. 4

[30] Q. Zhao, P. Tan, Q. Dai, L. Shen, E. Wu, and S. Lin. A closed-form solution to retinex with nonlocal texture constraints. *PAMI*, 34(7):1437–1444, 2012. 5

[31] D. Zoran, P. Isola, D. Krishnan, and W. T. Freeman. Learning ordinal relationships for mid-level vision. In *ICCV*, 2015. 2