Learning Lightness from Human Judgement on Relative Reflectance

Takuya Narihira
UC Berkeley / ICSI / Sony Corp.
takuya.narihira@berkeley.edu

Michael Maire
TTI Chicago
mmaire@ttic.edu

Stella X. Yu
UC Berkeley / ICSI
stellayu@berkeley.edu

Abstract

We develop a new approach to inferring lightness, the perceived reflectance of surfaces, from a single image. Classic methods view this problem from the perspective of intrinsic image decomposition, where an image is separated into reflectance and shading components. Rather than reason about reflectance and shading together, we learn to directly predict lightness differences between pixels.

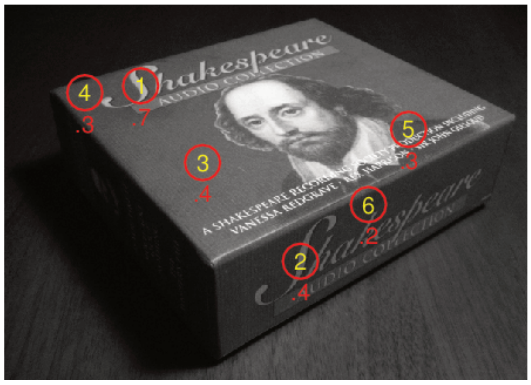
Large-scale training from human judgement data on relative reflectance, and patch representations built using deep networks, provide the foundation for our model. Benchmarked on the Intrinsic Images in the Wild dataset [4], our local lightness model achieves on-par performance with the state-of-the-art global lightness model, which incorporates multiple shading/reflectance priors and simultaneous reasoning between pairs of pixels in a dense conditional random field formulation.

1. Introduction

Human vision is remarkable at decoding the physical reflectance of an object despite variations of illumination cast upon it. This subjective constancy of *lightness*, which refers to the *perceived reflectance*, turns out to be hard to achieve computationally based entirely on the objective intensity of light recorded in an image [1, 25, 5, 13, 20, 11].

Figure 1 illustrates complicated ranking relationships between intensity and lightness on pairs of pixels [25]:

- *Different intensity* $\rightarrow$ *same lightness*: Letter S at circled locations 1 and 2 is seen to be the same white paint whether it is on the shaded face of the box; likewise the unevenly lit background gray at locations 3 and 4 is never mistaken for different paint colors.
- *Positive intensity differences* $\rightarrow$ *negative lightness differences*: The black clothes at location 5 are always correctly seen to be darker despite receiving in fact more light than the background gray at location 6.
- *Same intensity* $\rightarrow$ *different lightness*: Locations 2 and 3, 4 and 5, are of the same level of luminance but easily seen as painted with different shades of grays.



intensity: $I_1 > I_2 = I_3 > I_4 = I_5 > I_6$
lightness: $L_1 = L_2 > L_3 = L_4 = I_6 > L_5$

Figure 1. **Intensity vs lightness.** The subjective experience of lightness has a non-trivial relationship with the objective intensity of light in an image. This image features a 3D box whose top and front surfaces have identical text and background paint colors. Six pixels are labeled, and marked with their intensity values between 0 and 1. Their lightness, *i.e.* perceived reflectance, is ordered very differently from their measured intensity. Source: [25].

Conventional lightness modeling aims to recover the physical reflectance by teasing apart the confounded factors of illumination and reflectance from the image intensity, in the so-called *intrinsic image decomposition* [3] framework.

The basic idea is that, while the intensity at the pixels themselves is a poor indicator of lightness, their neighboring pixels often have enough variations for revealing the underlying illumination and reflectance, each of which has distinctive characteristics in the set of natural images [1]. The lightness results from identifying and discounting the factor of illumination from the intensity.

Specifically, the intrinsic image model assumes that the image intensity I is the product of reflectance image R and shading image S, and lightness L is simply the solution of R attempted by the visual system given its knowledge about

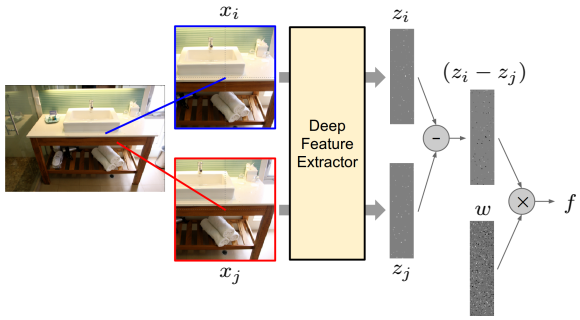


Figure 2. **Direct learning of pairwise lightness relationships.** Given two image patches x_i and x_j, we extract their deep features z_i and z_j, and train a classifier f that turns their difference $z_i - z_j$ into a pairwise lightness ordering.

the regularity of reflectance and shading:

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

$$L = R^*, \text{s.t. priors}(R, S) \tag{2}$$

Such a decomposition would be ill-defined without priors, and the general strategy is to exploit strong priors in order to constrain the search space for the solution (R^*, S^*) that satisfies the per-pixel factorization.

Many forms of priors reflecting the spatial and statistical regularity in a broader context have been explored. For example, reflectance is assumed piecewise constant [15, 16, 2], or from a sparse set [18, 9, 22], whereas shading is similar among nearby pixels [8], or more likely to take certain values than others [2, 16]. Good priors can also be learned generatively using deep belief nets [23]. The solution has been sought globally among all pairs of pixels through nonlocal texture constraints [26] or dense connected conditional random fields (CRFs) [4].

Our approach involves a complete change in intuition and strategy. We learn a lightness model directly from data, leveraging a training set of many relative reflectance comparisons made by human subjects. That is, we focus on learning the relative ordering of L, *i.e.* $L_i - L_j$, directly from contextual cues present in two local image patches, x_i and x_j, without resorting to an absolute pixel-wise decomposition of I into plausible R and S.

Formally, we learn the relative magnitude of lightness at locations i and j as a function of features z_i, z_j, extracted from local image patches, x_i, x_j, centered at i and j:

$$L_i - L_j = f(z_i, z_j) \tag{3}$$

Our intuition is that features extracted at increasingly larger neighborhoods of two pixels may capture the illumination and reflectance contexts. We need to know not what they are exactly, but whether and how they differ in order to render a judgement on the two pixels' lightness ordering.

Our model is built upon two key elements of recent work:

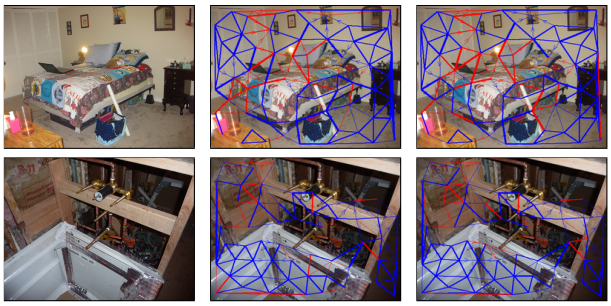


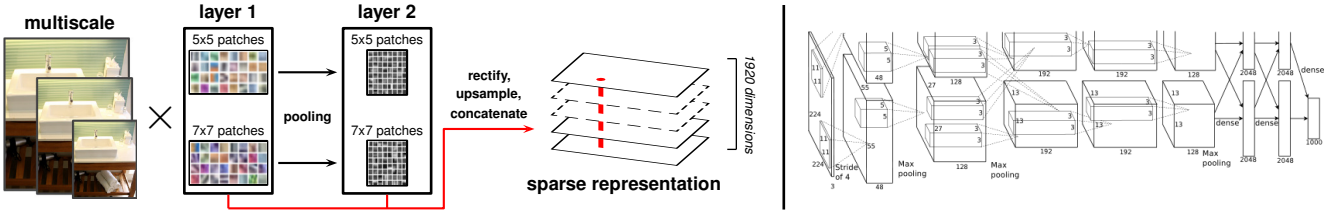
Figure 3. **Example pairwise lightness comparison results.** We show HSC and CNN classifier results for some IIW [4] test images. Arrows on graph edges point from brighter to darker regions according to the ground-truth; undirected edges indicate equal lightness. Edge thickness denotes human confidence, while edge color denotes classifier correctness (red for mistakes). Our errors tend to occur at challenging places such as low human confidence (thin red lines) and separate lightness contexts (thick red lines).

1. The Intrinsic Images in the Wild (IIW) dataset [4] provides a large collection of ground-truth in the form of human judgements of relative reflectance: 5230 indoor images with a total of 872,161 pairs of comparisons, about 106 ± 45 comparisons per image. Unlike the MIT intrinsic image dataset [24], where R and S are given as absolute ground-truth, these pairwise comparisons take three values: *same, lighter, or darker*, and they are noisy across human subjects.

2. Rich contextual features computed through either hierarchical sparse coding (HSC) [6, 17] or deep convolutional neural networks (CNN) [14, 12] provide an informative fine-to-coarse, small-to-large context feature vector representation of every patch in the input image, enabling a simpler and direct local classification approach without heavy reliance on any hand-designed global priors or expensive inference algorithms.

Our direct approach (Figure 2) is a departure from virtually all past methods which focus on differentiating these two aspects, reflectance and shading, either by learning to discriminate features and edges between these two aspects [24], or by imposing or discovering more priors that can be used to constrain each aspect [15, 16, 2, 18, 9, 22, 8, 26, 4]. Any hand-designed priors are completely absent from our algorithm; only implicit priors that are themselves learned from data have a role in our system.

Surprisingly, presented with a large collection of pairs of crude relative reflectance, we are able to learn a simple local linear classifier with lightness comparison accuracy on par with or better than the state-of-the-art model which relies on multiple priors and a dense CRF for global reasoning [4]. See sample results in Figure 3.



Hierarchical Sparse Coding [6, 17] **Convolutional Neural Network** [14]

Figure 4. **Deep feature extractors.** We consider two different procedures as implementations of the deep feature extractor in Figure 2. *Left:* We adapt the sparse coding strategy of [17], which builds on [6], and encode multiple image scales against multiple patch dictionaries. A second encoding layer operates on pooled output of the first. Concatenating sparse codes across layers at corresponding spatial locations yields per-pixel descriptors. Only a subset of the dictionaries are visualized here. *Right:* We utilize the convolutional neural network architecture of [14] with an image patch as input and treat the activations in the final layer as a spatially localized descriptor for the patch.

Our work is complementary to the global integration approach proposed in [25], where simple pairwise intensity differences across multiple scales are used as features and the global lightness ordering results from a reconciliation of all such pairwise cues in the entire image. While this model is demonstrated on challenging synthetic images, it is unclear how it can be applied to natural images with much more complex visual appearances.

Our work is the first to use complex deep features with a simple local classification rule for lightness prediction in natural images. It bypasses intrinsic image decomposition, yet could potentially be used by an intrinsic image model as a more informative local potential to improve *e.g.* CRF-based [4] or embedding-based [25] globalization approaches.

Section 2 describes in detail the two rich feature representations, HSC and CNN, that we consider for use as patch descriptors. Section 3 covers our learned model for pairwise lightness relationships, as well as the process of extracting, from this pairwise model, the linear classifier for direct construction of lightness maps. Section 4 provides experimental results and benchmarks, while Section 5 concludes.

2. Patch Representations

Motivated by their recent successes in other vision applications, we consider both hierarchical sparse coding [17, 6] and convolutional neural networks [14] for use as feature extractors in our system architecture. Figure 4 illustrates these approaches and we now briefly summarize each.

2.1. Hierarchical Sparse Coding

We borrow the patch representation strategy of [17], which in turn builds upon the work of [6]. As developed in [17], hierarchical sparse coding is an efficient algorithm for obtaining rich, high-dimensional, yet sparse, per-pixel feature descriptors. This sparseness translates into fast application of linear classifiers densely across the image.

Distinguishing characteristics as compared to CNN features include that HSC features are learned generatively and extracted from a multilayer slice of a deep network. Hence, they capture multiple different levels of abstraction. We review HSC here, but refer readers to [17] for more detail.

In a standard sparse coding setting, a patch $x \subset I$ is expressed as a sparse linear combination of at most K of N atoms from an overcomplete patch dictionary $D = [d_0, d_1, \ldots, d_{N-1}]$. Denoting by z the vector of combination coefficients, the encoding problem is:

$$\underset{z}{\operatorname{argmin}} ||x - Dz||^2 \quad s.t.\ ||z||_0 \leq K \tag{4}$$

While there is not a computationally efficient procedure for finding the exact optimal z, greedy approximation algorithms work well in practice [19, 21]. Having obtained z, we can simply treat it as a feature vector for x in the setting of classification or regression.

Encoding every patch in I against the same dictionary D yields a sparse N-dimensional coefficient vector for each pixel, or equivalently, a (sparse) image with N channels that lives on the same two-dimensional grid as I. In the hierarchical sparse coding setting, we treat this N-channel output from an initial layer of sparse coding as an input, after pooling, to another layer of sparse coding against a new dictionary of higher-dimensional atoms.

Our hierarchy of sparse coding layers is also "multipath" in the sense that at each layer, we also encode against several different dictionaries, for different patch sizes, and we encode the image at three different scales. Figure 4 illustrates this property. The feature vector representation z of an input patch is the concatenation of sparse coefficient vectors from all paths through the network. After concatenation, we rectify feature vectors:

$$z \leftarrow \left[\ \max(z^T, 0),\ \ \max(-z^T, 0)\ \right]^T \tag{5}$$

The dictionaries for each stage of encoding are learned in a generative manner, by sampling a collection of patches

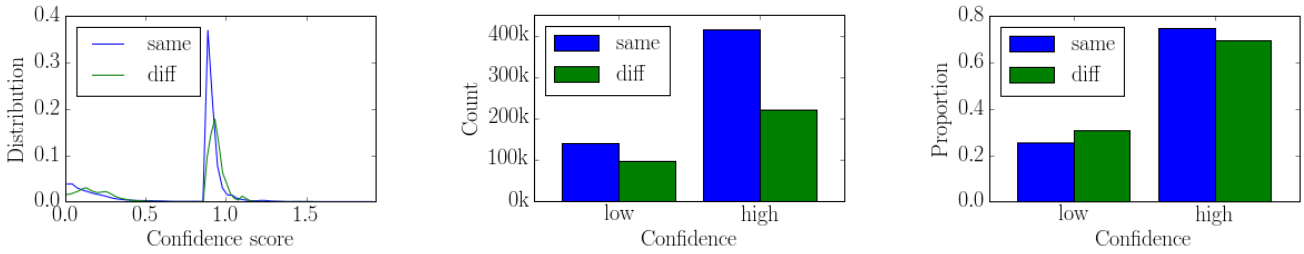


Figure 5. **Statistics of human reflectance annotations.** *Left:* On the IIW dataset [4], humans usually have high confidence in their relative reflectance judgements when annotating both same reflectance and lighter/darker pairs. *Middle:* Total count of high- and low-confidence ground-truth pairs of each reflectance type. *Right:* Proportion of ground-truth pairs of each type having high- and low-confidence.

$X = [x_0, x_1, ...]$ from training images (or the representation of training images output from the previous layer), and applying the MI-KSVD algorithm [6] to approximate a solution to:

$$\underset{D,\ Z}{\operatorname{argmin}} \left[||X - DZ||_F^2 + \lambda \sum_{i=0}^{N-1} \sum_{j=0, j\neq i}^{N-1} |d_i^T d_j| \right] \quad (6)$$
$$s.t.\ \forall i,\ ||d_i||_2 = 1 \text{ and } \forall n,\ ||z_n||_0 \leq K$$

where $||\cdot||_F$ denotes the Frobenius norm and K is the specified sparsity level.

In our experiments, we learn dictionaries of $N = 32$ and 64 atoms for each of 5×5, 7×7, and 9×9 patches in the first layer. The second layer uses dictionaries of 64 atoms for 5×5 and 7×7 patches of the 32-dimensional 5×5 first layer output. Concatenation and rectification produce a 1920-dimensional feature vector at every pixel in the image.

2.2. Convolutional Neural Network

We use the Caffe [12] implementation of the 7-layer convolutional neural network of [14] and take the 4096-dimensional activations of the final fully connected layer as a feature descriptor for a patch presented as input to the network. Each patch is taken from a 227×227 window surrounding a location (vertex) in a labeled pair (edge) from the IIW dataset. Note that the entire network can be trained via backpropagation in a discriminative manner, in contrast to HSC, which learns dictionary weights generatively. We attempt to train the CNN from both randomly initialized weights and weights initialized by pre-training on ImageNet [7].

3. Lightness Classifier

Humans are capable of making reliable comparisons of the reflectance at two different points in the same scene [4]. We leverage this observation in building an automatic classifier that replicates human relative lightness judgements.

Following Equation (3) and choosing a linear form for f with either the HSC or CNN representations of the preceding section serving as features z, we have

$$f(z_i, z_j) = w^T(z_i - z_j) \quad (7)$$

We learn HSC classifier weights w by ridge ranking regression on the human ground-truth data for reflectance:

$$\min \varepsilon(w) = \sum_{i,j} \log\left(1 + \exp(-J_{ij} w^T(z_i - z_j))\right) + \gamma w^T w \quad (8)$$

where:

$$J_{ij} = \begin{cases} 1, & R_i^h > R_j^h \\ -1, & R_i^h < R_j^h \end{cases} \quad (9)$$

$$C_{ij} = \text{ confidence in } J_{ij} \quad (10)$$

Here R^h refers to human ratings of relative reflectance (lightness) on the IIW dataset. γ calibrates regularization. For each example where humans judge equal reflectance ($R_i^h = R_j^h$), we create two virtual examples with both $R_{ij} = 1$ and $R_{ij} = -1$ in order to force prediction $f(z_i, z_j)$ toward zero. Although confidence C_{ij} is present in the dataset, we do not use it for training. To train our CNN model, we use the standard cross entropy classification loss.

Figure 5 provides some insight into the distribution of human confidence (C_{ij}) on their pairwise reflectance ratings. While the IIW dataset contains more examples of point pairs with the same reflectance than different reflectance (as judged by humans), human annotators tend to be confident when reporting either relationship. The slight imbalance in the overall number of training examples of each type is thus not problematic.

We can equivalently regard w as either a linear predictor for relative lightness from a difference of feature descriptors, or as a lightness potential function acting on the feature representation at point. Define *lightness potential*:

$$g(z) = w^T z \quad (11)$$

Then our relative lightness classifier is a difference:

$$f(z_i, z_j) = w^T(z_i - z_j) = g(z_i) - g(z_j) \quad (12)$$

HSC pairwise predictions **HSC lightness potential**

Figure 6. **Global lightness rank.** From a linear difference model trained for relative lightness judgements, we extract a per-pixel lightness potential (a global lightness ranking). *Left:* Pairwise results of our HSC model of relative lightness between pixels located at the graph nodes. Arrows point from brighter to darker according to model predictions. *Right:* Densely predicted relative lightness potential using HSC features.

This perspective is useful as we can evaluate lightness potential $g(\cdot)$ locally at every pixel to recover a global rank ordering of the lightness of all pixels in the image.

In the case of HSC descriptors, such evaluation is cheap since z is extremely sparse. The HSC implementation of [17] generates descriptors densely over the image. Figure 6 shows example potentials obtained with HSC features. They are similar to a scaled version of the reflectance channel in the traditional intrinsic image decomposition.

While dense evaluation of $g(\cdot)$ is also possible using CNN features, the CNN implementation we use [12] is not targeted to fully convolutional evaluation over an arbitrary-sized input and it is prohibitively expensive to run independently for all patches in an image. We therefore focus on the task of matching human judgements for query patch pairs.

As used in the IIW dataset benchmark, given reflectance R at two points, a discrete judgement in terms of *lighter, darker, same* is rendered based on the reflectance ratio test:

$$\widehat{J}_{ij}(R;\delta) = \begin{cases} 1, & \text{if } \frac{R_i}{R_j} > 1+\delta \\ -1, & \text{if } \frac{R_j}{R_i} > 1+\delta \\ 0, & \text{otherwise} \end{cases} \tag{13}$$

where δ is a threshold below which relative reflectance changes are considered insignificant.

Our linear classifier can be interpreted as performing the reflectance ratio test according to a difference test in the transformed log reflectance domain:

$$g(z) = w^T z = \alpha \log R \tag{14}$$

$$\begin{aligned} f(z_i, z_j) &= g(z_i) - g(z_j) = \alpha \log R_i - \alpha \log R_j \\ &= \alpha \log \frac{R_i}{R_j}. \end{aligned} \tag{15}$$

Note that $f(z_i, z_j)$ does not change its sign, 0 or ± 1, no matter what α is, and it can be considered the internal lightness difference between two points; whereas scaling parameter α controls the rate at which perceived reflectance relates to the lightness potential, and it can be considered a sensitivity parameter that turns an absolute internal difference into an external *perceivable lightness difference*.

Now given predefined threshold δ for the reflectance ratio test, we tune α so that $\widehat{J}(f;\alpha,\delta)$ for perceivable lightness differences best matches human judgements J on the training set:

$$\widehat{J}_{ij}(f;\alpha,\delta) = \begin{cases} 1, & \text{if } f(z_i, z_j) > \alpha \log(1+\delta) \\ -1, & \text{if } f(z_j, z_i) > \alpha \log(1+\delta) \\ 0, & \text{otherwise} \end{cases} \tag{16}$$

Note that α (and thus the decision threshold $\alpha \log(1+\delta)$) is optimized over the entire training set, not per image or per pair of query points.

Previous work on intrinsic image decomposition [10, 4] is not similarly tuned specifically for the pairwise lightness judgement task. While their reflectance is directly modeled and does not need an interpretation like ours, to ensure a fair benchmark comparison, we take their output reflectance maps and report performance at optimal δ for their algorithms (an equivalent form of rescaling).

4. Results

We use the Intrinsic Images in the Wild (IIW) dataset [4] for both training and testing our lightness classifier. There are 44 ± 16 query points and 106 ± 45 query pairs between these points per image (they are the nodes and edges in Figure 3), over a total of 5230 images. Each query to a human subject was in the form of "which point has a darker surface color?". Bell *et al*. [4] obtained a total of 4,880,372 responses from 1381 Amazon Mechanical Turk workers and then aggregated this individual human pairwise judgement data into 875,833 comparisons across 5230 photos, each with a confidence measuring how consistent the result is among workers.

This large set of pairwise comparisons has been used to benchmark several reflectance models [4]. The weighted human disagreement rate (WHDR) is proposed to measure the percent of human judgements that an algorithm disagrees with, weighted by the confidence C_{ij} of each judge-

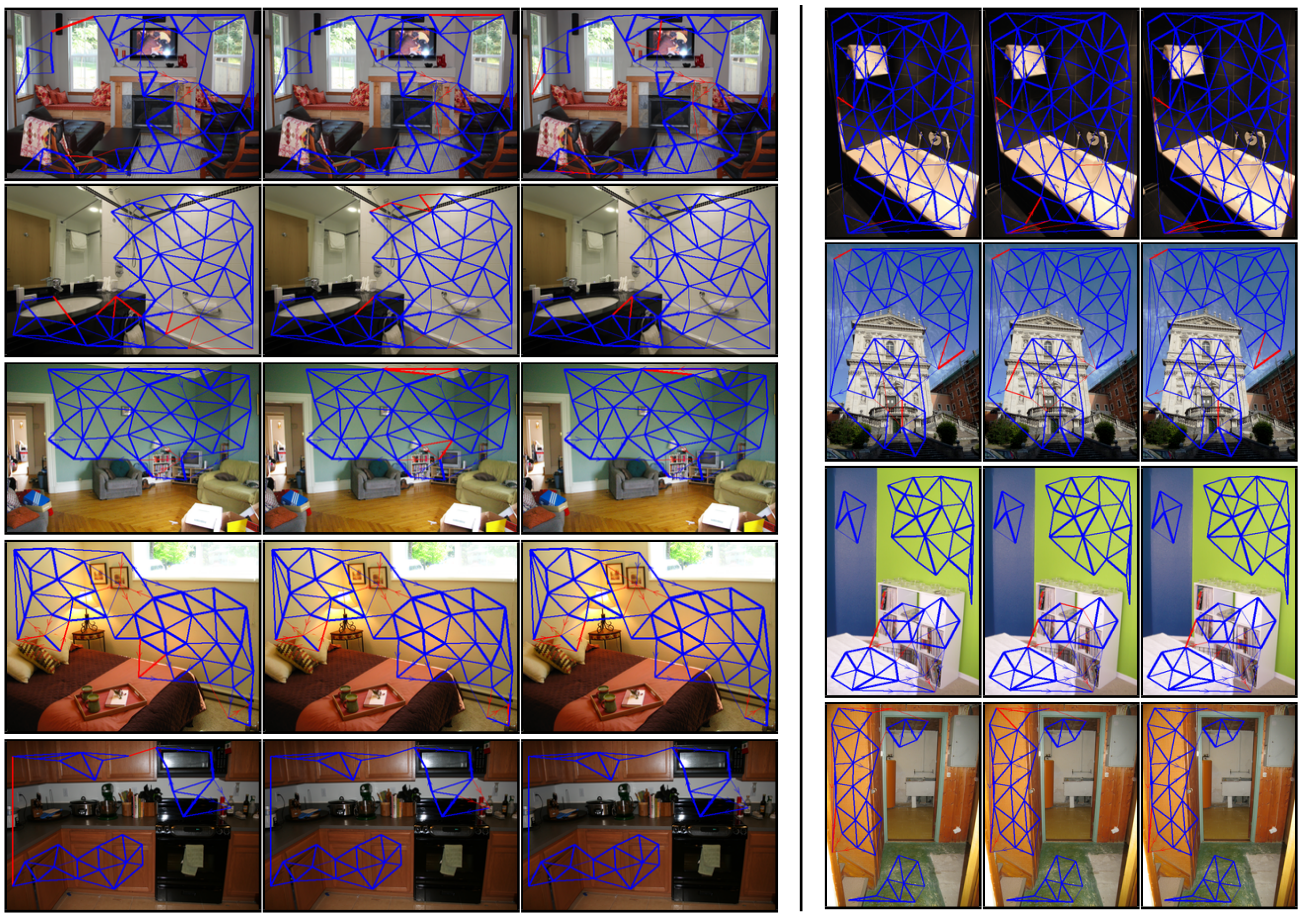


Figure 7. **Good lightness predictions.** We visualize the benchmarked pairwise lightness predictions of both versions of our model (HSC and CNN) and the CRF model of [4]. As in Figure 3, arrows on graph edges point according to ground-truth (towards darker endpoint) and color denotes classifier correctness (red for mistakes). Here, we show examples where all models perform well.

ment on point pair (i, j):

$$WHDR_\delta(J, R) = \frac{\sum_{ij} C_{ij}(J_{ij} \neq \widehat{J}_{ij}(R; \delta))}{\sum_{ij} C_{ij}} \quad (17)$$

where J_{ij} and $\widehat{J}_{ij}$ are human ground-truth and machine predictions, respectively. Note that such a score is computed for each image, and then all the scores are averaged across photos in the dataset to yield an overall benchmark number for an algorithm.

The state-of-the-art result is by Bell *et al.* [4], with a dense CRF model on intrinsic image decomposition incorporating many forms of reflectance and shading priors explored in the literature. Note that human subjects are not necessarily consistent with each other, or between all the query pairs per individual. Overall human consistency is at WHDR of 7.5% [4].

We introduce an additional metric, classification error rate $(1 - ACC)$, to measure an algorithm's performance per query edge, instead of the mean accuracy (or error by WHDR) over all the query edges per image, as the number of query edges vary greatly between images. We define the classification accuracy (ACC) as:

$$ACC_\delta(J, R; \delta) = \text{mean}(J_{ij} = \widehat{J}_{ij}(R; \delta), \quad \forall i, j) \quad (18)$$

For our experiments, we split the IIW dataset into 80% training examples and 20% testing examples as follows. We sort the dataset by image ID, then take every 5 examples in order and use the first one as testing and the rest for training. We re-evaluated several approaches reported in [4] on this split to make sure that scores matched those reported in [4] and the particular training-test split had negligible influence.

Table 1 compares our results with the state-of-the-art CRF approach as well as other competing methods. Our models are trained on a fixed 80% of data and with no further optimization. Table 1 shows that, despite the simplicity

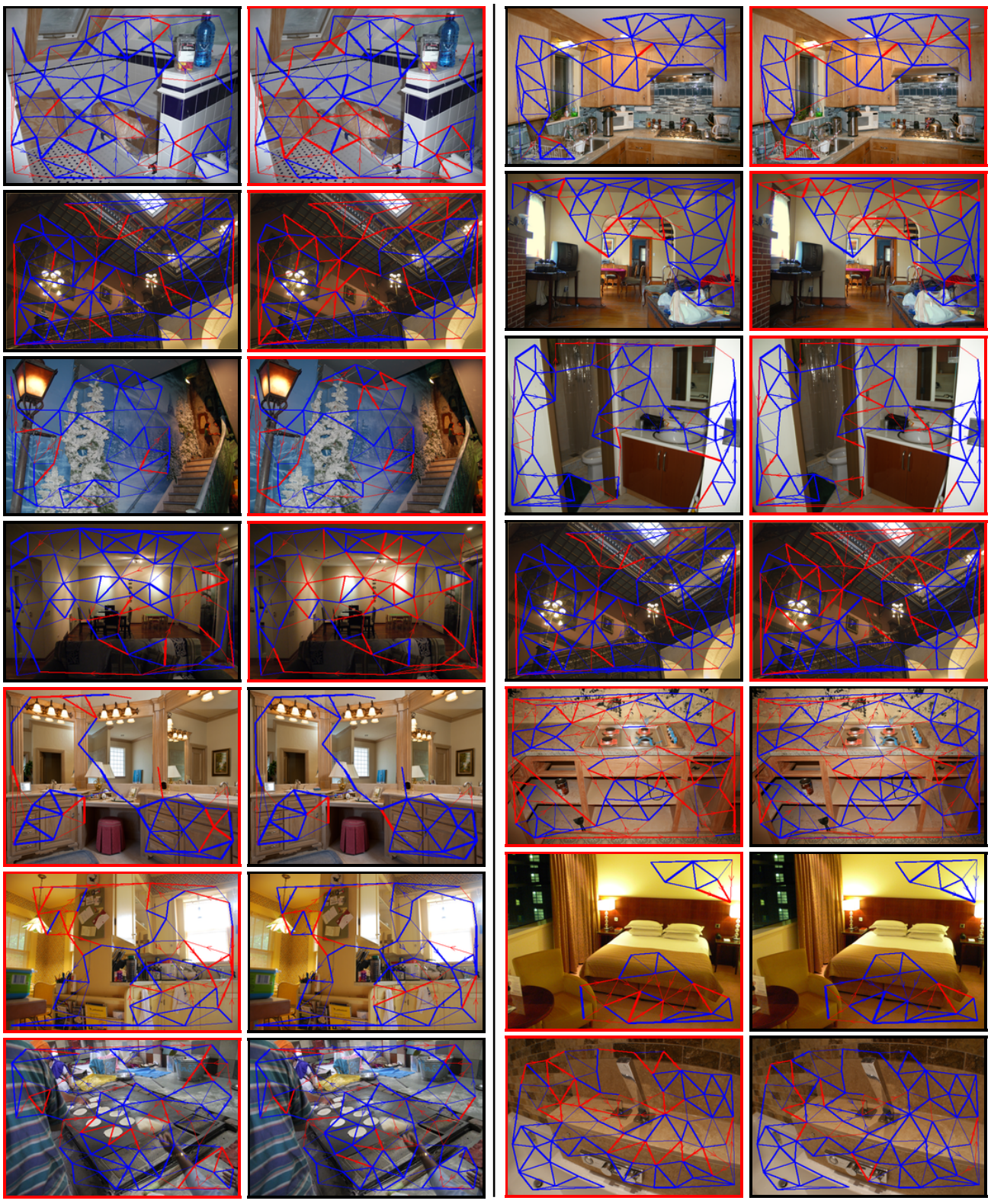


Figure 8. **Mistakes in lightness predictions.** As in Figure 7, we visualize predictions of our model (HSC and CNN versions) and the CRF model of [4]. Here we show examples where there is a large performance gap between them. *Left:* Examples with significant performance gap between CRF and HSC models. *Right:* Examples with significant performance gap between CRF and CNN models. For each pair, results of the poorly performing model are highlighted with a red border around the image displaying its predictions.

	WHDR (%)	Error Rate (%)
Ours (HSC)	20.9	24.5
Ours (CNN)	18.3	22.3
Ours (CNN-ImageNet)	**18.1**	**22.0**
CRF [4] (rescaled)	18.6	22.3
Retinex-Color [10] (rescaled)	19.5	23.3
Retinex-Gray [10] (rescaled)	19.8	23.8
Shen and Yeo [22] (rescaled)	23.2	26.1
Zhao *et al*. [26] (rescaled)	22.8	26.4
CRF [4]	20.6	25.6
Retinex-Color [10]	26.9	32.4
Retinex-Gray [10]	26.8	32.3
Shen and Yeo [22]	32.5	35.1
Zhao *et al*. [26]	23.8	28.2

Table 1. **Intrinsic Images in the Wild benchmark results.** For each algorithm, we display the weighted human disagreement rate (WHDR, lower is better), as well as the error rate on classifying the sign of lightness change between pairs of points labeled in the ground-truth. We include our own re-evaluation of competing methods, which closely matches the performance reported in [4]. In addition, we report performance of a rescaled version of competing methods, which specifically optimizes their output for the pairwise classification task. Our algorithm is on par with the CRF approach developed by [4] for state-of-the-art performance. We refer the reader to [4] for comparison to an expanded set of prior work.

of our local classifier model, our CNN performance is on par with that of the best method on the test set, and ahead of all others. Here, CNN refers to CNN trained from randomly initialized weights while CNN-ImageNet is initialized with pre-trained weights on ImageNet. We do not see significant performance differences between them.

Figure 7 shows sample images where our method performs equally well as the CRF approach. The query edges in these images tend to involve points in roughly uniform patches, where the judgement tends to be less ambiguous and easier. Figure 8 shows sample images where one technique is far superior to another. Our model and the CRF approach have different failure modes.

Our method with HSC features seems to make more mistakes than CRF for points in low-light and high-light conditions, and the types of mistakes are characteristic: it tends to mistake different lightness (edges with arrows) as same lightness in low light, and mistake same lightness (edges without arrows) as different lightness in high light. Our method with CNN features provides a more uniform improvement over CRF across lighting conditions. Figure 9 provides additional visualization of classifier performance as a function of local lighting (mean intensity) and texture (local variance in pixel intensity) properties of patches.

The CRF method seems to make more mistakes than ours in textured areas or an image with many small structures. This error pattern can be explained by the reflectance

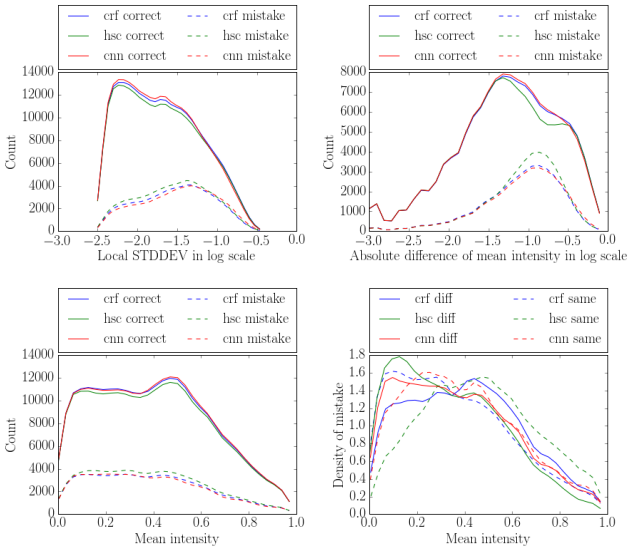


Figure 9. **Mistakes as a function of local patch statistics.** *Top Left:* Plotting errors against patch intensity variance reveals that the the CNN model has lower mistake count in smooth areas (low pixel intensity variance). *Top Right:* The HSC model makes more mistakes when comparing two patches of drastically differing intensities. *Bottom Left:* We compare model errors with respect to the mean intensity of two patches. Our CNN model is uniformly better than others over the range of patch mean intensity. *Bottom Right:* We organize mistakes by type (ground-truth *different* or *same* lightness patch pairs incorrectly classified). HSC makes more mistakes on ground-truth *diff* pairs in low-light, and more mistakes on *same* pairs in high-light conditions.

sparsity (about 20 reflectances per image) and piecewise constancy priors used by the CRF model. When the image contains many different reflectances, *e.g*. each small structure assumes a different reflectance, the CRF gives inaccurate decomposition results due to the assumption violation, and hence more errors.

5. Conclusion

We demonstrate a local classification model that performs as well as far more complicated schemes on the task of recovering relative lightness. Use of rich patch representations, obtained via hierarchical sparse coding or convolutional neural networks, and a large amount of training data, enable us to learn a better local model than was previously possible. Such models open up new areas of exploration on the classic problem of intrinsic image decomposition.

Acknowledgments. We thank Ayan Chakrabarti for valuable discussion.

References

[1] E. H. Adelson. Lightness perception and lightness illusions. In M. Gazzaniga, editor, *The cognitive neurosciences*, pages 339–51. MIT Press, Cambridge, MA, 1999.

[2] J. T. Barron and J. Malik. Intrinsic scene properties from a single RGB-D image. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

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

[4] S. Bell, K. Bala, and N. Sanvely. Intrinsic images in the wild. In *ACM Trans. on Graphics*, 2014.

[5] B. Blakeslee, W. Pasieka, and M. E. McCourt. Oriented multiscale spatial filtering and constrast normalization: a parsimonious model of brightness induction in a continuum of stimuli including white, howe and simulteneous brightness contrast. *Vision Research*, (45):607–15, 2005.

[6] L. Bo, X. Ren, and D. Fox. Multipath sparse coding using hierarchical matching pursuit. *CVPR*, 2013.

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[8] E. Garces, A. Munoz, J. Lopez-Moreno, and D. Gutierrez. Intrinsic images by clustering. In *Computer Graphics Forum (Eurographics Symposium on Rendering)*, 2012.

[9] P. Gehler, C. Roth, M. Kiefel, L. Zhang, and B. Scholkopf. Recovering intrinsic images with a global sparsity prior on reflectance. In *Neural Information Processing Systems*, 2011.

[10] R. Grosse, M. K. Johnson, E. H. Adelson, and W. T. Freeman. Ground truth dataset and baseline evaluations for intrinsic image algorithms. In *International Conference on Computer Vision*, 2009.

[11] B. Horn. Determining lightness from an image. *Computer Graphics and Image Processing*, 3:277–99, 1974.

[12] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *arXiv preprint arXiv:1408.5093*, 2014.

[13] F. Kelly and S. Grossberg. Neural dynamics of 3-d surface perception: figure-ground separation and lightness perception. *Perception and Psychophysics*, (62):1596–618, 2000.

[14] A. Krizhevsky, S.Ilya, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Neural Information Processing Systems*, 2012.

[15] E. H. Land and J. J. McCann. Lightness and retinex theory. *Journal of Optical Society of America*, 61(1):1–11, 1971.

[16] Z. Liao, J. Rock, Y. Wang, and D. Forsyth. Non-parametric filtering for geometric detail extraction and material representation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[17] M. Maire, S. X. Yu, and P. Perona. Reconstructive sparse code transfer for contour detection and semantic labeling. *ACCV*, 2014.

[18] I. Omer and M. Werman. Color lines: image specific color representation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2004.

[19] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. *Asilomar Conference on Signals, Systems and Computers*, 1993.

[20] W. D. Ross and L. Pessoa. Lightness from contrast: a selective integration model. *Perception and Psychophysics*, (62):1160–81, 2000.

[21] R. Rubinstein, M. Zibulevsky, and M. Elad. Efficient implementation of the K-SVD algorithm using batch orthogonal matching pursuit. 2008.

[22] L. Shen and C. Yeo. Intrinsic images decomposition using a local and global sparse representation of reflectance. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[23] Y. Tang, R. Salakhutdinov, and G. Hinton. Deep lambertian networks. In *International Conference on Machine Learning*, 2012.

[24] M. Tappen, W. Freeman, and E. Adelson. Recovering intrinsic images from a single image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2005.

[25] S. X. Yu. Angular embedding: from jarring intensity differences to perceived luminance. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2302–9, 2009.

[26] Q. Zhao, P. Tan, Q. Dai, L. Shen, E. Wu, and S. Lin. A closed-form solution to retinex with nonlocal texture constraints. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2012.