Coherent Intrinsic Images from Photo Collections

Pierre-Yves Laffont[1] Adrien Bousseau[1] Sylvain Paris[2] Frédo Durand[3] George Drettakis[1]

[1] REVES / INRIA Sophia Antipolis [2] Adobe [3] MIT CSAIL

Input photo-collection Coherent reflectance Individual illuminations

Figure 1: *Our method leverages the heterogeneity of photo collections to automatically decompose photographs of a scene into reflectance and illumination layers. The extracted reflectance layers are coherent across all views, while the illumination captures the shading and shadow variations proper to each picture. Here we show the decomposition of three photos in the collection.*

Abstract

An intrinsic image is a decomposition of a photo into an illumination layer and a reflectance layer, which enables powerful editing such as the alteration of an object's material independently of its illumination. However, decomposing a single photo is highly under-constrained and existing methods require user assistance or handle only simple scenes. In this paper, we compute intrinsic decompositions using several images of the same scene under different viewpoints and lighting conditions. We use multi-view stereo to automatically reconstruct 3D points and normals from which we derive relationships between reflectance values at different locations, across multiple views and consequently different lighting conditions. We use robust estimation to reliably identify reflectance ratios between pairs of points. From these, we infer constraints for our optimization and enforce a coherent solution across multiple views and illuminations. Our results demonstrate that this constrained optimization yields high-quality and coherent intrinsic decompositions of complex scenes. We illustrate how these decompositions can be used for image-based illumination transfer and transitions between views with consistent lighting.

Keywords: intrinsic images, photo collections

Links: DL PDF

1 Introduction

Image collections aggregate many images of a scene from a variety of viewpoints and are often captured under different illuminations. The variation of illumination in a collection has often been seen as a nuisance that is distracting during navigation or, at best an interesting source of visual diversity. Inspired by existing work on time-lapse sequences [Weiss 2001; Matsushita et al. 2004], we consider these variations as a rich source of information to compute *intrinsic images*, i.e., to decompose photos into the product of an illumination layer by a reflectance layer [Barrow and Tenenbaum 1978]. This decomposition is an ill-posed problem since an infinity of reflectance and illumination configurations can produce the same image, and so far automatic techniques are limited to simple objects [Grosse et al. 2009], while real-world scenes require user assistance [Bousseau et al. 2009], detailed geometry [Troccoli and Allen 2008; Haber et al. 2009], or varying illumination with a fixed or restricted viewpoint [Weiss 2001; Liu et al. 2008].

In this paper, we exploit the rich information provided by multiple viewpoints and illuminations in an image collection to process complex scenes without user assistance, nor precise and complete geometry. Furthermore, we enforce that the decomposition be *coherent*, which means that the reflectance of a scene point should be the same in all images.

The observation of a point under different unknown illuminations does not help directly with the fundamental ambiguity of intrinsic images. Any triplet R, G, B is a possible reflectance solution for which the illumination of the point in each image is its pixel value divided by R, G, B. We overcome this difficulty by processing pairs of points. We consider the ratio of radiance between two points, which is equal to the ratio of reflectance if the points share the same illumination. A contribution of this paper is to identify pairs of points that are likely to have similar illumination across most conditions. For this, we leverage sparse 3D information from multi-view stereo as well as a simple statistical criterion on the distribution of the observed ratios. These ratios give us a set of equations relating the reflectance of pairs of sparse scene points, and consequently of sparse pixels where the scene points project in the input images. To infer the reflectance and illumination for all the pixels, we build on image-guided propagation [Levin et al. 2008; Bousseau et al. 2009]. We augment it with a term to force the estimated reflectance of a given 3D point to be the same in all the images in which it is visible. This yields a large sparse linear system, which we solve in an interleaved manner. By enforcing coherence in the reflectance layer we obtain a common "reflectance space" for all input views, while we extract the color variations proper to each image in the illumination layer.

Our automatic estimation of coherent intrinsic image decompositions from photo collections relies on the following contributions:

- A method to robustly identify reliable reflectance constraints between pairs of pixels, based on multi-view stereo and a statistical criterion.

- An optimization approach which uses the constraints within and across images to perform an intrinsic image decomposition with *coherent* reflectance in all views of a scene.

We run our method on 9 different scenes, including a synthetic benchmark with ground truth values, which allows for a comparison to several previous methods. We use our intrinsic images for image-based illumination transfer between photographs captured from different viewpoints. Our coherent reflectance layers enable stable transitions between views by applying a single illumination condition to all images.

2 Related Work

Single-Image Methods. Retinex [Horn 1986] distinguishes gradient illumination based on magnitude, which was extended by Tappen et al. [2005] using machine learning. Shen et al. [2008] and Zhao et al. [2012] assume that similar texture implies the same reflectance. In contrast, Shen and Yeo [2011] assume that similar chromaticity indicates same reflectance for neighboring pixels and that each image only contains a small number of reflectances. These methods work well with isolated objects [Grosse et al. 2009]. Bousseau et al. [2009] and Shen et al. [2011] require user annotations, whereas we need an automatic method to handle the large number of images in a collection.

Multiple-Images Methods. For timelapse sequences, Weiss [2001] applies a median operator in the gradient domain as a robust estimator of the reflectance derivatives. However, Matsushita et al. [2004] observe that this estimator produces poor decompositions when neighboring pixels have different normals and the input images do not cover the illumination directions uniformly. They instead use the median estimator to detect flat surfaces on which they enforce smooth illumination. Sunkavalli et al. [2007] use timelapse sequences to derive a shadow mask and images lit only by the sky or the sun. Matusik et al. [2004] additionally capture light probes to estimate a reflectance field. Our approach builds on this family of work, but we seek to handle images captured from multiple viewpoints, and avoid the sometimes cumbersome timelapse capture process.

Inverse rendering, e.g., [Yu and Malik 1998; Yu et al. 1999; Debevec and et al. 2004] requires detailed geometric models and challenging non-linear fitting. Troccoli and Allen [2008] use a laser scan and multiple lighting and viewing conditions to perform relighting and estimate Lambertian reflectance. In addition to a detailed geometry, they rely on a user-assisted shadow detector. Haber et al. [2009] estimate BRDFs and distant illumination in 3D scenes reconstructed with multi-view stereo. However, as stated by the authors, manual intervention remains necessary to correct the geometry and ensure accurate visibility computation for shadow removal. In contrast, our work relies on statistical analysis and image-guided propagation to automatically estimate reflectance from incomplete 3D reconstructions, even when shadow casters are not observed in the input photographs. While our method assumes Lambertian reflectance, it produces pixel-accurate decompositions that are well suited for image editing and image-based rendering. In contrast, to obtain pixel-accurate results, model-based approaches typically require high-precision laser-scans [Debevec and et al. 2004], rather than the less accurate multi-view stereo 3D reconstructions as used e.g., in [Haber et al. 2009]. Laffont et al. [2012] use a light probe and multiple images under a *single* lighting condition to reconstruct a sparse geometric representation similar to ours to constrain the intrinsic image decomposition. Their approach requires the same lighting in all views, which is not the case in photo collections. In work developed concurrently, Lee et al. [2012] use a depth camera to compute intrinsic decompositions of video sequences. They constrain the decomposition in a way similar to our approach, using surface orientation and temporal coherence between frames. However, they target indoor scenes with dense 3D reconstruction, while we deal with photo-collections taken under varying lighting conditions and with sparse 3D reconstructions.

Photo Collections. Photo-sharing websites such as Flickr© and Picasa© contain millions of photographs of famous landmarks captured under different viewpoints and illumination conditions. Photo collections of less famous places are also becoming available thanks to initiatives like the collaborative game PhotoCity [Tuite et al. 2011]. The wide availability of photos on the internet has been exploited for many computer graphics applications including scene completion [Hays and Efros 2007] and virtual tourism [Snavely et al. 2006; Snavely et al. 2008]. Liu et al. [2008] extend Weiss's algorithm to colorize grayscale photographs from photo collections. They use a homography or a mesh-based warping to project images on a single viewpoint. This is well adapted to images viewed from similar directions, but tends to produce blurry decompositions in the presence of large viewpoint changes.

Finally, Garg et al. [2009] apply dimensionality reduction on photo collections to estimate representative basis images that span the space of appearance of a scene. While some of the basis images model illumination effects, this "blind" decomposition does not extract a single reflectance and illumination pair for each input image.

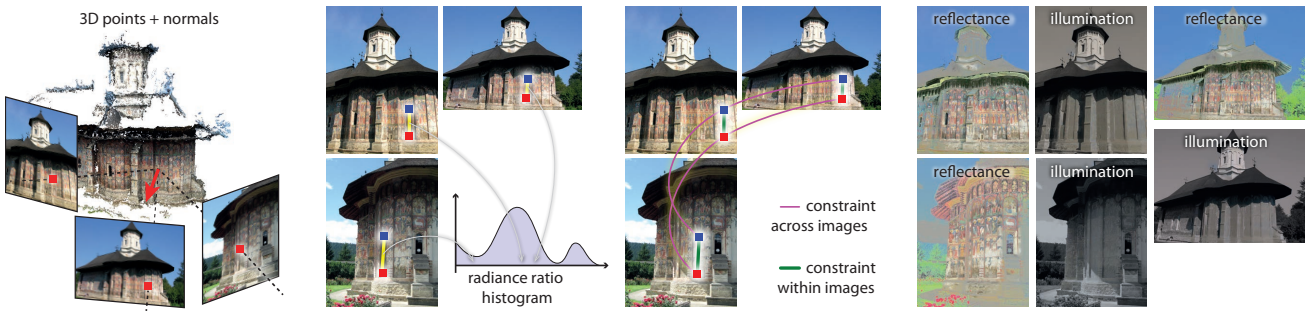
3 Overview

We take as input a collection of photographs $\{\mathbf{I}_i\}$ of a given scene captured from different viewpoints and under varying illumination. We seek to decompose each input image into an illumination layer $\mathbf{S}_i$ and a reflectance layer $\mathbf{R}_i$ so that, for each pixel p and each color channel c, $\mathbf{I}_{ic}(p) = \mathbf{S}_{ic}(p)\mathbf{R}_{ic}(p)$. Furthermore, whereas the illumination is expected to change from image to image, we assume that the scene is mostly Lambertian so that the reflectance of a point is constant across images. In the following, we drop the color channel subscript c and assume per-channel operations, unless stated explicitly.

In order to leverage the multiple illumination conditions, we need to relate scene points in different images. For this, we apply standard multiview-stereo [Furukawa and Ponce 2009] (Fig. 2(a)), which produces an oriented point cloud of the scene and estimates for each point the list of images where it appears. For ease of notation, we make 3D projection implicit and denote the value of the pixel where point $\mathbf{p}$ projects in image i as $\mathbf{I}_i(\mathbf{p})$.

We next infer ratios of reflectance between pairs of 3D points (Fig. 2(b)). For a pair of points ($\mathbf{p}$, $\mathbf{q}$), we consider the distribution of ratios of pixel radiance $\mathbf{I}_i(\mathbf{p}) \,/\, \mathbf{I}_i(\mathbf{q})$ in all the images where both points are visible. The ratio of reflectance is equal to the median ratio of radiance if the two points have the same illumination in most lighting conditions. A contribution of our work is to identify pairs of points that share the same illumination based on geometric criteria and on the distribution of radiance ratios.

Our last step solves for the illumination layer at each image based on a linear least squares formulation (Fig. 2(c)). It includes the constraints on reflectance ratios (depicted as green edges in Fig. 2(c)), an image-guided interpolation inspired by Levin et al. [2008] and Bousseau et al. [2009], and terms that force reflectance to be the same in all images (edges in magenta).



(a) off-the-shelf multiview stereo (b) reflectance ratio inference (c) least-squares propagation within and across images (d) output illumination and reflectance

Figure 2: *Our method infers reflectance ratios between points of a scene and then expresses the computation of illumination in all images in a unified least-square optimization system.*

4 Reflectance ratios

Our method relies on reflectance ratios inferred from the multiple illumination conditions. In order to relate points in different images, we reconstruct a sparse set of 3D points and normals, and introduce a statistical criterion to reliably infer reflectance ratios.

4.1 Relations on reflectance between pairs of points

If two points $\mathbf{p}$ and $\mathbf{q}$ have the same normal $\vec{n}$ and receive the same incoming radiance, then the variations of the observed radiances $\mathbf{I}$ are only due to the variations of the scene reflectance $\mathbf{R}$.

Assuming Lambertian surfaces, the radiance $\mathbf{I}$ towards the camera at each non-emissive point $\mathbf{p}$ is given by the following equation:

$$\mathbf{I}(\mathbf{p}) \;=\; \mathbf{R}(\mathbf{p}) \int_{\Omega} \mathbf{L}(\mathbf{p},\vec{\omega})\; (-\vec{\omega}\cdot\vec{n}(\mathbf{p}))\; d\vec{\omega} \qquad (1)$$

where $\mathbf{L}(\mathbf{p},\vec{\omega})$ is the incoming radiance arriving at $\mathbf{p}$ from direction $\vec{\omega}$, $\vec{n}(\mathbf{p})$ is the normal at $\mathbf{p}$, and Ω is the hemisphere centered at $\vec{n}(\mathbf{p})$.

Given a pair of points $\mathbf{p}$ and $\mathbf{q}$ with the same normal $\vec{n}$, we can express the ratio of radiance between the two points as

$$\frac{\mathbf{I}(\mathbf{q})}{\mathbf{I}(\mathbf{p})} = \frac{\mathbf{R}(\mathbf{q})}{\mathbf{R}(\mathbf{p})}\;\frac{\int_{\Omega}\mathbf{L}(\mathbf{q},\vec{\omega})\;(-\vec{\omega}\cdot\vec{n})\;d\vec{\omega}}{\int_{\Omega}\mathbf{L}(\mathbf{p},\vec{\omega})\;(-\vec{\omega}\cdot\vec{n})\;d\vec{\omega}}. \qquad (2)$$

If the incoming radiance $\mathbf{L}$ is identical for both points, then the ratio of reflectances $\mathbf{R}(\mathbf{q})\;/\;\mathbf{R}(\mathbf{p})$ is equal to the ratio of radiances $\mathbf{I}(\mathbf{q})\;/\;\mathbf{I}(\mathbf{p})$. From multiview stereo we have a normal estimate for each point, and it is straightforward to find points with similar normals. We next find an image where lighting conditions at $\mathbf{p}$ and $\mathbf{q}$ match. For points $\mathbf{p}$ and $\mathbf{q}$ which are close, the likelihood that a shadow boundary falls between them is low. Thus for most images in which these points are visible, the radiance ratio is equal to the reflectance ratio. However, lighting may still not match in a few images. Inspired by the work of Weiss [2001] and Matsushita et al. [2004] in the context of timelapse sequences, we use the median operator as a robust estimator to deal with such rare cases:

$$\frac{\mathbf{R}(\mathbf{q})}{\mathbf{R}(\mathbf{p})} = \underset{i\in\mathcal{I}(\mathbf{p},\mathbf{q})}{\text{median}}\left(\frac{\mathbf{I}_i(\mathbf{q})}{\mathbf{I}_i(\mathbf{p})}\right) \qquad (3)$$

where the median is taken only over the images of the set $\mathcal{I}(\mathbf{p},\mathbf{q}) \subset \{\mathbf{I}_i\}$ in which both $\mathbf{p}$ and $\mathbf{q}$ are visible.

Ambient occlusion. Our derivation so far assumes that the illumination depends only on the normal orientation and is independent of the location. However, for scenes with strong concavities, differences in visibility might cause two points with similar normals to have different illumination on average, because one of them might be in shadow more often. We compensate for this by evaluating the ambient occlusion factor $\alpha(\mathbf{p})$, that is, the proportion of the hemisphere visible from $\mathbf{p}$. We compute ambient occlusion by casting rays from the 3D points in the upper hemisphere around the normal, and intersecting them with a geometry proxy created with standard Poisson mesh reconstruction [1]. The estimation of ambient occlusion is robust to inaccurate geometry, since it averages the contribution of incoming light from all directions of the hemisphere. For points in the shadow, Eq. 2 becomes:

$$\frac{\mathbf{I}(\mathbf{q})}{\mathbf{I}(\mathbf{p})} \;=\; \frac{\mathbf{R}(\mathbf{q})}{\mathbf{R}(\mathbf{p})}\,\frac{\alpha(\mathbf{q})}{\alpha(\mathbf{p})} \qquad (4)$$

We account for this by multiplying the ratio $\mathbf{I}(\mathbf{q})/\mathbf{I}(\mathbf{p})$ by $\alpha(\mathbf{p})/\alpha(\mathbf{q})$ to correct the reflectance ratio estimated in Eq. 3.

4.2 Selection of constrained pairs

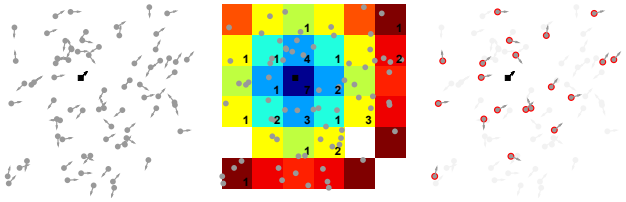
Given the set of 3D points, we need to select a tractable number of pairs whose median ratio is likely to be a good estimate of the reflectance ratio. Based on the above discussion, we first selectively subsample the set of all possible constraints according to geometric factors, i.e., normals and distance. We then discard unreliable constraints with a simple statistical criterion on the observed ratios.

Geometric criterion. For each 3D point, we select a set of candidate pairs that follow the geometric assumptions in Sec. 4.1. In most cases, the two points of a pair should be nearby and have similar normals. However, we also wish to obtain a well-connected graph of constraints, with a few pairs consisting of points which are further apart or with varying orientations. Our approach consists in sampling candidate pairs by controlling the distribution of their spatial extent and orientation discrepancy. Note that this step only selects *candidate pairs*, on which constraints *might* be applied; unreliable pairs will be discarded in the next step of the algorithm.

We define the distance $d_{\vec{n}}$ on normal orientation between two points $\mathbf{p}$ and $\mathbf{q}$ from the dot product between their normals:

$$d_{\vec{n}}(\mathbf{p},\mathbf{q}) = |1 - \vec{n}(\mathbf{p})\cdot\vec{n}(\mathbf{q})|\,. \qquad (5)$$

[1] In practice, we use the Poisson reconstruction in MeshLab (http://meshlab.sourceforge.net)



(a) Initial point cloud (b) Number of samples per cell (color: d_{3D} to reference cell) (c) Final sampled points

Figure 3: *2D Illustration of our sampling algorithm for a single point. (a) Given an oriented point cloud, we wish to select N points so that their distances d_{3D} and $d_{\vec{n}}$ to a reference point (black square) follow normal distributions. (b) We first embed the point cloud in a grid and compute Euclidean distances to the cell containing the reference point; the distance is color-coded from blue to red. We infer a sampling probability for each cell based on d_{3D} as described in Algorithm 1, from which we draw N samples to choose the number of points to select in each cell, shown as black numbers. (c) Finally, we sample the corresponding number of points within each cell based on the normal discrepancy $d_{\vec{n}}$. Note that a point can be sampled multiple times if its cell contains too few points.*

We set $d_{3D}(\mathbf{p}, \mathbf{q})$ to be the Euclidean 3D distance, representing the spatial proximity of two points.

Our goal is to select N candidate pairs of points so that $d_{\vec{n}}$ and d_{3D} follow normal distributions $\mathcal{N}(\sigma_{\vec{n}})$ and $\mathcal{N}(\sigma_{3D})$. The parameter $\sigma_{\vec{n}}$ accounts for surfaces with low curvature and inaccuracy in the normals estimated from multiview stereo. We set σ_{3D} to 20% of the spatial extent of each scene, and $\sigma_{\vec{n}} = 0.3$ for all our results.

For a given point $\mathbf{p}$, we sample the density functions in two steps. First we select a subset of points according to $\mathcal{N}(\sigma_{3D})$, and then we sample this subset according to $\mathcal{N}(\sigma_{\vec{n}})$. In both cases, the major difficulty resides in properly accounting for the non-uniform distribution of the distance $d \in \{d_{\vec{n}}, d_{3D}\}$ in the point cloud generated by multiview stereo. We account for these non-uniform distributions with the following algorithm:

Algorithm 1 Sampling according to 3D distances or normals

1. Estimate the density of distances $f_{\text{original}}(d(\mathbf{p}, \mathbf{q}))$ of all points $\mathbf{q}$ to the current point $\mathbf{p}$. We use the Matlab `ksdensity` function, which computes a probability density estimate of distances to $\mathbf{p}$ from a set of samples $d(\mathbf{p}, \mathbf{q})$ by accumulating normal kernel functions centered on each sample.
2. Assign to each point $\mathbf{q}$ a sampling probability based on desired distribution $\mathcal{N}(\sigma)$ and the density of distances f_{original}:
$$\Pr(\mathbf{q}) \;=\; \exp\left(-\frac{d(\mathbf{p},\mathbf{q})^2}{2\sigma^2}\right) / f_{\text{original}}(d(\mathbf{p},\mathbf{q}))$$
3. Select a subset of points according to their probabilities $\Pr(\mathbf{q})$ using inversion sampling.

In practice, we accelerate the sampling of $\mathcal{N}(\sigma_{3D})$ by first embedding the point cloud in a 3D grid (with 10^3 non-empty cells on average). We then apply Algorithm 1 to the grid cells instead of the points, ignoring empty cells and computing d_{3D} at the cell centers. As a result of this first sampling we obtain a list of cells and the number of points that we need to choose in each cell to obtain a total of N pairs. We then apply Algorithm 1 according to $d_{\vec{n}}$, with the caveat that we only consider points from the cells that should be sampled, and we apply inversion sampling independently in each cell to select the proper number of points. We illustrate this process in Fig. 3 and supplemental materials, and provide Matlab code[2].

[2]https://www-sop.inria.fr/reves/Basilic/2012/LBPDD12/

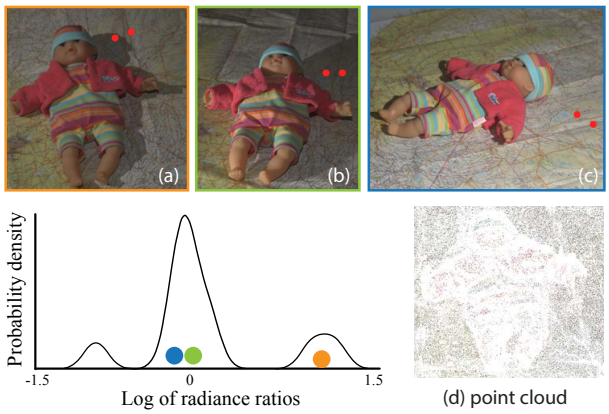


Figure 4: *Analysis of the distribution of radiance ratio (red channel, log scale) between two 3D points (red dots) with similar normals, under varying viewpoints and lighting. The PDF has a dominant lobe, corresponding to (b) and (c) where both points receive approximately the same incoming radiance. In (a), the light is visible from only one of the points and the corresponding radiance ratio falls in a side lobe. (d) shows the point cloud for image (a).*

Our sampling strategy ensures a good distribution of pairs of points, with many "short distance" pairs around the point and a few "longer distance" pairs. We also experimented with a simple threshold that selects the pairs with the highest score based on $d_{\vec{n}}$ and d_{3D}, but this naive strategy tends to only select short distance pairs with identical normals, yielding a weakly connected graph of constraints. This results in isolated regions in the final optimization. We used 30 candidate pairs per point in all our examples, and keep at most 1.5 million candidate pairs per scene.

Cuboid scenes are seemingly problematic since pairwise constraints cannot connect orthogonal faces. However, the faces may be indirectly connected via other objects in the scene. The solution is also influenced by a smoothness prior (Sec. 5.2) and a coherence term (Sec. 5.3). In our experiments, these additional constraints were enough to obtain plausible decompositions even on cuboid scenes (Fig. 7, left; Fig. 9, bottom row).

Photometric statistical criterion. Each candidate pair $(\mathbf{p}, \mathbf{q})$ can be observed in a subset of input images $\mathcal{I}(\mathbf{p}, \mathbf{q})$. Figure 4 illustrates the probability density function (PDF) of the ratio of radiances of a pair over multiple images with varying lighting. When the two points fulfill our assumptions, the distribution has a dominant lobe well captured by the median operator. In such a case, the reflectance ratio of the pair can be estimated with the median. However when the two points receive different incoming radiance in more than 50% of the images, the distribution is spread and not necessarily centered at the median. We detect and reject such unreliable pairs, by counting the observations of the radiance ratio that are far from the median value. The observation of pair $(\mathbf{p}, \mathbf{q})$ in image j is considered far from the median if

$$\left|\log\left(\frac{\mathbf{I}_j(\mathbf{q})}{\mathbf{I}_j(\mathbf{p})}\right) - \underset{i\in\mathcal{I}(\mathbf{p},\mathbf{q})}{\text{median}} \log\left(\frac{\mathbf{I}_i(\mathbf{q})}{\mathbf{I}_i(\mathbf{p})}\right)\right| > 0.15 \qquad (6)$$

in at least one channel. We consider a pair to be unreliable if it has less than 50% of the radiance ratio values close to the median, or if it is visible in less than 5 images (too few observations). Candidate pairs that are considered reliable will be used to constrain the intrinsic image decomposition (Sec. 5.1).

5 Multi-Image Guided Decomposition

We now have a sparse set of constraints on the ratio of reflectance at 3D points. To obtain values everywhere, we formulate an energy function over the RGB illumination $\mathbf{S}$ at each pixel of each image. Our energy includes data terms on the reflectance ratios, an image-guided interpolation term, and a set of constraints that enforce the coherence of the reflectance between multiple images. This results in a large sparse linear least square system, which we solve in a staggered fashion.

5.1 Pairwise reflectance constraints

Given the ratio between the reflectances of pixels corresponding to points $\mathbf{p}$ and $\mathbf{q}$ in Eq. 3, we deduce ratio $\mathbf{Q}_j(\mathbf{p},\mathbf{q})$ between the illumination of the corresponding pixels in image j:

$$\mathbf{Q}_j(\mathbf{p},\mathbf{q}) = \frac{\mathbf{S}_j(\mathbf{p})}{\mathbf{S}_j(\mathbf{q})} = \frac{\mathbf{I}_j(\mathbf{p})}{\mathbf{I}_j(\mathbf{q})}\frac{\mathbf{R}(\mathbf{q})}{\mathbf{R}(\mathbf{p})} \tag{7a}$$

$$= \frac{\mathbf{I}_j(\mathbf{p})}{\mathbf{I}_j(\mathbf{q})} \underset{i\in\mathcal{I}(\mathbf{p},\mathbf{q})}{\text{median}} \left(\frac{\mathbf{I}_i(\mathbf{q})}{\mathbf{I}_i(\mathbf{p})}\right) \tag{7b}$$

where $\mathbf{S}_j$ is the illumination layer of image j. This equation lets us write a constraint on the unknown illumination values:

$$\mathbf{Q}_j(\mathbf{p},\mathbf{q})^{\frac{1}{2}}\,\mathbf{S}_j(\mathbf{q}) \;=\; \mathbf{Q}_j(\mathbf{p},\mathbf{q})^{-\frac{1}{2}}\,\mathbf{S}_j(\mathbf{p}) \tag{8}$$

We combine the contribution of all the constrained pairs selected in Sec. 4.2 in all the images where they are visible, and express these constraints in a least-squares sense to get the energy $E_{\text{constraints}}$:

$$\sum_j \sum_{(\mathbf{p},\mathbf{q})} \left[\mathbf{Q}_j(\mathbf{p},\mathbf{q})^{\frac{1}{2}}\,\mathbf{S}_j(\mathbf{q}) - \mathbf{Q}_j(\mathbf{p},\mathbf{q})^{-\frac{1}{2}}\,\mathbf{S}_j(\mathbf{p})\right]^2 \tag{9}$$

In practice, we have one such term for each RGB channel.

5.2 Smoothness

We build our smoothness prior on the intrinsic images algorithm of Bousseau et al. [2009] that was designed to propagate sparse user indications for separating reflectance and illumination in a single image, and on the closely related *Matting Laplacian* introduced by Levin et al. [2008] for scribble-based matting. The former assumes a linear relationship between the unknowns and the image channels and the latter an affine relationship. We experimented with both, and while the intrinsic image prior captures variations of illumination at a long distance from the constrained pixels, we show in the supplemental materials that the matting prior yields smoother illumination in regions with varying reflectance, especially in our context where many pixels are constrained.

The matting prior translates into a local energy for each pixel neighborhood that relates the color at a pixel x with the illumination value in each channel $\mathbf{S}_{jc}(x)$ using an affine model:

$$\sum_{c\in\{r,g,b\}} \sum_{y\in\mathcal{W}_j^x} \left(\mathbf{S}_{jc}(y) - \mathbf{a}_{jc}^x \cdot \mathbf{I}_j(y) - b_{jc}^x\right)^2 + \epsilon\,(\mathbf{a}_{jc}^x)^2 \tag{10}$$

where $\mathcal{W}_j^x$ is a 3×3 window centered on x, $\mathbf{a}_j^x$ and b_j^x are the unknown parameters of the affine model, constant over the window, and $\epsilon = 10^{-6}$ is a parameter controlling the regularization $(\mathbf{a}_j^x)^2$ that favors smooth solutions. Levin et al. [2008] showed that $\mathbf{a}_j^x$ and b_j^x can be expressed as functions of $\mathbf{S}_j$ and removed from the system. Then, summing over all pixels and all images yields an energy that only depends on the illumination, and can be expressed in matrix form:

$$E_{\text{smoothness}} = \sum_{c\in\{r,g,b\}} \sum_j \hat{\mathbf{S}}_{jc}^{\mathsf{T}} M_{jc} \hat{\mathbf{S}}_{jc} \tag{11}$$

where the vectors $\hat{\mathbf{S}}_j$ stack the unknown illumination values in image j and the matrices M_j encode the smoothness prior over each pixel neighborhood in this image (see the paper by Levin et al. [2008] for the complete derivation).

We found that it is beneficial to add a grayscale regularization for scenes with small concavities in shadow. Because these areas often have no (or very few) reconstructed 3D points, they are influenced by their surrounding lit areas and illumination tends to be overestimated. For such scenes, we add the term below to favor illumination values close to the image luminance:

$$\sum_x \sum_{c\in\{\text{r,g,b}\}} \left(\mathbf{S}_{jc}(x) - \frac{1}{3}\left[\mathbf{I}_{j\text{r}}(x) + \mathbf{I}_{j\text{g}}(x) + \mathbf{I}_{j\text{b}}(x)\right]\right)^2 \tag{12}$$

We use a small weight (10^{-3}) so that this term affects only regions with no other constraints. We show in supplemental material that although results are satisfying without it, this term helps further improve the decomposition.

5.3 Coherent reflectance

For photo collections, it is important to ensure that the intrinsic image decomposition is coherent across different views. We impose additional constraints across images by enforcing the reflectance of a 3D point to be constant over all views where it appears.

Consider the case where a given point $\mathbf{p}$ is visible in two images $\mathbf{I}_m$ and $\mathbf{I}_n$. For each such pair (m,n) of images we want to force the pixels corresponding to $\mathbf{p}$ to have the same reflectance, and thus infer a constraint on their illumination:

$$\begin{aligned}\mathbf{R}_m(\mathbf{p}) = \mathbf{R}_n(\mathbf{p}) \Rightarrow \frac{\mathbf{I}_m(\mathbf{p})}{\mathbf{S}_m(\mathbf{p})} &= \frac{\mathbf{I}_n(\mathbf{p})}{\mathbf{S}_n(\mathbf{p})} \\ \Rightarrow \mathbf{I}_m(\mathbf{p})\,\mathbf{S}_n(\mathbf{p}) &= \mathbf{I}_n(\mathbf{p})\,\mathbf{S}_m(\mathbf{p})\end{aligned} \tag{13}$$

We denote $\mathcal{I}(\mathbf{p}) \subset \{\mathbf{I}_i\}$ the subset of images where the point $\mathbf{p}$ is visible. Summing the contribution of every pair of images where a point appears gives us an additional energy term $E_{\text{coherence}}$ that encourages coherent reflectance across images:

$$\sum_{\mathbf{p}} \sum_{m\in\mathcal{I}(\mathbf{p})} \sum_{\substack{n\in\mathcal{I}(\mathbf{p})\\ n>m}} \left(\mathbf{I}_m(\mathbf{p})\,\mathbf{S}_n(\mathbf{p}) - \mathbf{I}_n(\mathbf{p})\,\mathbf{S}_m(\mathbf{p})\right)^2 \tag{14}$$

This term generates a large number of constraints. We found that applying them only at the points selected in Sec. 4.2 yields equivalent results while reducing the complexity of the system. In addition, we describe an efficient solver in Sec. 5.4.

5.4 Solving the system

We combine the energy terms defined above with weights $w_{\text{constraints}} = 1$, $w_{\text{smoothness}} = 1$ and $w_{\text{coherence}} = 10$, fixed for all our results. Minimizing this global energy translates into solving a sparse linear system where the unknowns are the illumination values at each pixel of each image. We obtain the reflectance at each pixel by dividing the input images by the estimated illuminations.

Our system is large because it includes unknowns for all the pixels of all the images to decompose. To make things tractable, we use an

iterative approach akin to a blockwise Gauss-Seidel solver, where each iteration solves for the illumination of one image with the values in all the other images fixed. The advantage of this approach is that we can reduce Eq. 14 to a single term per point $\mathbf{p}$. To show this, we first write the energy $E^k_{\text{coherence}}(m, \mathbf{p})$ for point $\mathbf{p}$ in image m at iteration k:

$$\sum_{\substack{n\in\mathcal{I}(\mathbf{p})\\ n<m}} \left(\mathbf{I}_m(\mathbf{p})\,\mathbf{S}^k_n(\mathbf{p}) - \mathbf{I}_n(\mathbf{p})\,\mathbf{S}^k_m(\mathbf{p})\right)^2 + \sum_{\substack{n\in\mathcal{I}(\mathbf{p})\\ n>m}} \left(\mathbf{I}_m(\mathbf{p})\,\mathbf{S}^{(k-1)}_n(\mathbf{p}) - \mathbf{I}_n(\mathbf{p})\,\mathbf{S}^k_m(\mathbf{p})\right)^2 \quad (15)$$

In this energy, the only variable is $\mathbf{S}^k_m(\mathbf{p})$, everything else is fixed. Since all the terms in Eq. 15 are quadratic functions depending on the same variables, the energy can be rewritten as a single least-squares term, plus a constant which does not depend on $\mathbf{S}^k_m(\mathbf{p})$:

$$\left(\sum_{\substack{n\in\mathcal{I}\\ n\neq m}} \mathbf{I}_n^2\right)\left(\mathbf{S}^k_m - \frac{\mathbf{I}_m\left(\sum_{\substack{n\in\mathcal{I}\\ n\neq m}} \mathbf{I}_n\,\mathbf{S}^{\tilde{k}}_n\right)}{\sum_{\substack{n\in\mathcal{I}\\ n\neq m}} \mathbf{I}_n^2}\right)^2 + \text{constant} \quad (16)$$

where for clarity, we use the notation $\mathbf{S}^{\tilde{k}}_n = \mathbf{S}^k_n$ when $n < m$ and $\mathbf{S}^{(k+1)}_n$ when $n > m$, and omit the dependency on $\mathbf{p}$.

Eq. 16 expresses the inter-images constraints on $\mathbf{S}^k_m(\mathbf{p})$ as a single least-squares term, which shows that these constraints are tractable even though there is a quadratic number of them. Further, when we derive this term to obtain the corresponding linear equation used in our solver, the left factor and the denominator cancel out, ensuring that our system does not become unstable with small values of $\sum_{\substack{n\in\mathcal{I}\\ n\neq m}} \mathbf{I}_n^2$.

To initialize this iterative optimization, we compute an initial guess of the illumination in each image with an optimization where we only use the single-image terms $E_{\text{constraints}}$, $E_{\text{smoothness}}$, and the grayscale regularization. The energy decreases quickly during the first few iterations of the optimization process, then converges to a plateau value. We applied 4 iterations for all the results in this paper. Intermediate results after each iteration are shown as supplemental material.

6 Implementation and Results

3D Reconstruction. We first apply bundle adjustment [Wu et al. 2011] to estimate the parameters of the cameras and patch-based multi-view stereo [Furukawa and Ponce 2009] to generate a 3D point cloud of the scene. For each point, this algorithm also estimates the list of photographs where it appears. We compute normals over this point cloud using the PCA approach of Hoppe et al. [1992]. We used 103 images per scene on average to perform reconstruction, but this varies significantly depending on the scene (e.g., we used 11 image to reconstruct the "Doll" scene).

Point cloud resampling. Multi-view reconstruction processes full-sized photographs, while we apply our decomposition on smaller images for efficiency. Multiple nearby 3D points may project to the same pixels on the resized images. We downsample the point cloud so that at most one 3D point projects to each pixel in each image, using a greedy algorithm which gives priority to points that are visible in most images. To do so, we visit every pixel of every image, creating the point set as we proceed. At a given pixel, we first test if a point of the set already projects to it. If not, we choose the point which is visible in the largest number of images, and we add it to the set. We finally limit the size of the point cloud to $200k$ points. We also discard points that project on strong edges because their radiance tends to result from a mixture of reflectances that varies among images: we discard a point if the variance of the radiance in adjacent pixels is greater than 4×10^{-3}.

Performance. The average running time of our method is 90 minutes for the 9 scenes in this paper. Our unoptimized Matlab implementation of the sampling algorithm (Sec. 4.2) takes 52 min. on average and the selection of reliable constraints takes less than a minute. Each iteration of the optimization takes 6 min on average; we use Matlab's backslash operator to solve for each image within one iteration. We could greatly speed up our method by parallelizing the sampling of candidate pairs for each 3D point.

6.1 Intrinsic Decompositions

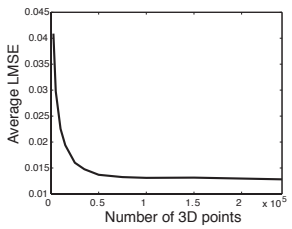
We demonstrate our method on three types of data. First we apply our method to synthetic data which allows a comparison to ground truth. We then show results of our method for captured scenes in which we have placed cameras and lights around objects in a room. We finally apply our method to online photo collections.

Evaluation on a Synthetic Scene. We evaluate our method against a ground truth decomposition that we rendered from a synthetic scene. We use a diffuse model of the St. Basil cathedral because it contains complex geometric details and a colorful spatially varying reflectance, in addition to occluded areas that are challenging for our approach (Sec. 4.1). We render the scene and compute ground truth illumination using path-tracing in PBRT [Pharr and Humphreys 2010], and obtain ground truth reflectance by dividing the rendering by the illumination. We use a physically-based sun and sky model [Preetham et al. 1999] for daylight, and captured environment maps for sunset/sunrise and night conditions. We rendered 30 different viewpoints over the course of three days (in summer, autumn and winter). To apply our method, we sample the 3D model to generate a 3D point cloud; this allows us to evaluate the performance of our algorithm independently of the quality of multi-view reconstruction. We provide the decompositions of 6 views as supplemental material, as well as all input and ground truth data.

Fig. 5 provides a visual comparison of our method against ground truth, as well as state-of-the-art automatic and user-assisted methods, all kindly provided by the authors of the previous work. In supplemental materials, we provide more images and comparisons to additional methods, including [Garces et al. 2012] and our implementation of [Weiss 2001] extended to multiview, which is inspired by [Liu et al. 2008]. In Fig. 6 we plot the Local Mean Squared Error (LMSE, as in [Grosse et al. 2009]) of each method with respect to ground truth, averaged over all views.

For this benchmark, our approach produces results that closely match ground truth and outperform single image methods. In particular, we successfully decompose the night picture while automatic methods fail to handle the yellow spot lights and blue shadows. Our method extracts most of colored texture from the illumination in this challenging case. Our method also produces coherent reflectance between all views, despite the drastic change of lighting.

We study the robustness of our algorithm by varying the number of 3D points in the point cloud, as shown in the inset graph on the right. Our approach still outperforms the best



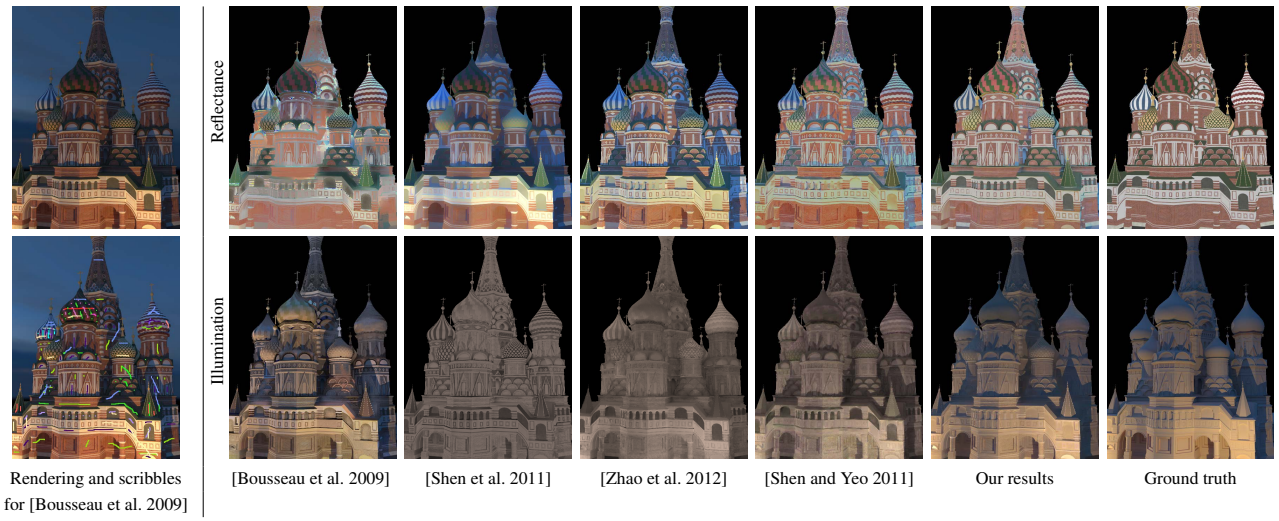


Figure 5: *Comparison to existing methods and ground truth on a synthetic rendering, generated with path tracing (see text for details). Reflectance and illumination images have been scaled to best match ground truth; sky pixels have been removed.*

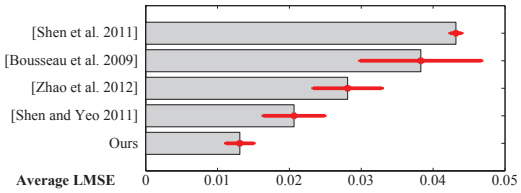


Figure 6: *Numerical evaluation of five intrinsic decomposition methods. Gray bars indicate Local Mean Squared Error averaged over the three comparison images, while red bars illustrate the standard deviation of LMSE across images.*

single-image based technique when only 15000 points are used. We also reconstruct a point cloud with PMVS, after specifying ground truth camera parameters since structure from motion techniques fail on our synthetic images. Our decomposition using this reconstruction yields an average LMSE of 0.01564, still significantly lower than all the approaches compared. Please see supplemental materials for the corresponding images.

Captured Scenes. We set up two indoors scenes containing small objects and used two light sources: a camera-mounted flash with low intensity, which simulates ambient light in shadows, while a remote-controlled flash produces strong lighting from a separate direction. This setup allows us to validate our algorithm on real photographs, while avoiding the difficulty inherent in internet photo collections, such as the use of different camera settings or contrast and color manipulation that affect the validity of our assumptions.

Fig. 7 shows our decomposition for the "Doll" and "Temple" scenes. We used 11 and 10 viewpoints respectively, and 7 different lighting conditions. Both scenes contain colored reflectances (cloth of the baby doll, texture of the tabletop) and strong hard shadows that are successfully decomposed by our method. Non-lambertian components of the reflectance (such as the specularities on the tablecloth) are assigned to the illumination layer, since coherency constraints enforce similar reflectance across images. We provide as supplemental material a visual comparison between our method and previous work on a similar "Doll" scene.

Figure 7: *Results of our decomposition on scenes captured with a flash. Note that the colored residual in the doll illumination is due mainly to indirect light.*

Internet Photo Collections. The last set of results we show is on internet photo collections of famous landmarks; we chose challenging scenes with interesting lighting and shadowing effects. We download images from Flickr© or Photosynth©, avoiding pictures that have been overly edited. We use 45 images on average to compute the pairwise reflectance ratios (Eq. 3), and perform the intrinsic decomposition on around 10 images per dataset. Table 1 lists the number of images used for each scene.

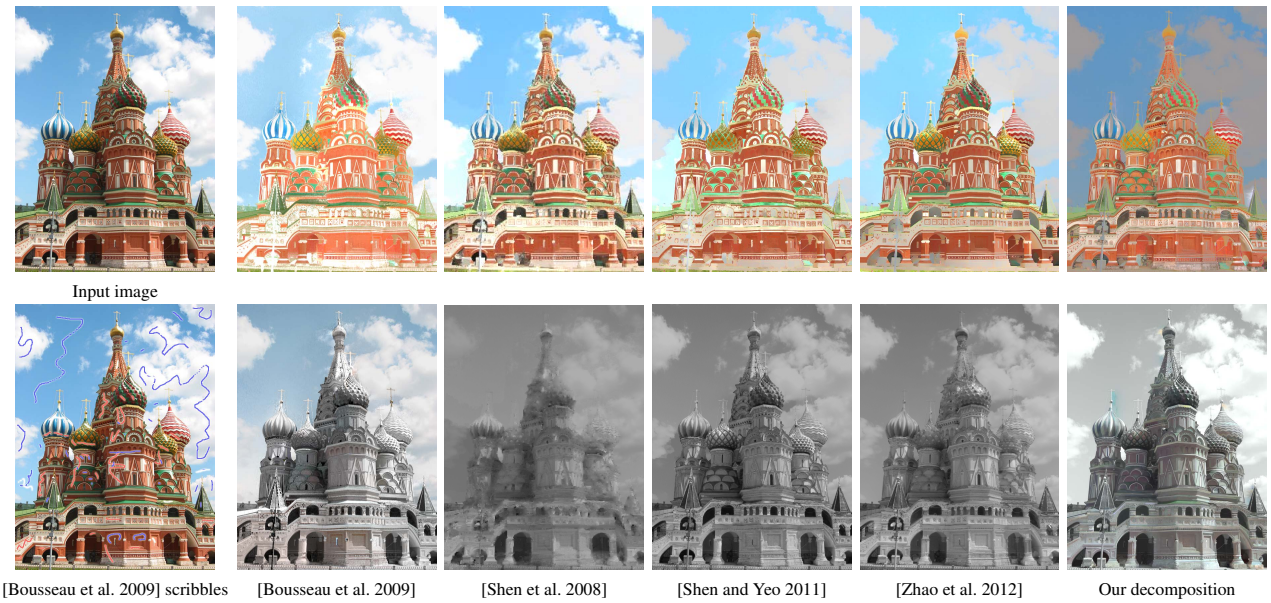


Figure 8: *Comparison between our approach and existing single-image methods on a picture from an online collection.*

We correct radial distortion using camera parameters estimated from scene reconstruction. We assume a gamma correction of 2.2 which is common for jpeg images. However, noise and non-linearities in the camera response can generate unreliable pixels which have very low values in some channels; in such cases we recover reliable information from other channels when available.

Fig. 9 illustrates our results on several scenes, namely St. Basil, Knossos and RizziHaus. Our method successfully decomposes the input image sets into intrinsic images, despite the complex spatially-varying reflectance. In Fig. 8 we present a side-by-side comparison with four existing single-image methods on a real picture of the St. Basil cathedral. We provide coherent reflectance (compare our result to Fig. 9, top row), which was not in the scope of single-image approaches. Our reflectance result is comparable in quality to the best previous work. Coherent reflectance results in some residual color in shading, although these residual are attenuated in other views (see Fig. 9, top row); this is discussed in Sec. 6.2.

In Fig. 1, our algorithm successfully disambiguates the complex texture on the lower facade where sparse 3D information is available. However, our decomposition assigns a similar grey reflectance to the steeple and roof of the monastery because very few 3D points are reconstructed in these areas. Without 3D points, the decomposition lacks pairwise and coherence constraints and relies mostly on the smoothness prior. Many single-image methods assume that pixels with similar chrominance share similar reflectance, which is likely to produce the same greyish reflectance as ours on this image, as shown in the supplemental materials.

Figure 9: *Results of our method on internet photocollections. Top: another view of the StBasil scene. The reflectance we extract is coherent with the one shown in Fig. 8. Bottom: the specular objects which cast shadows on the façade are a challenging case for multi-view stereo. Our method is able to extract their shadows despite the lack of a complete and accurate 3D reconstruction.*

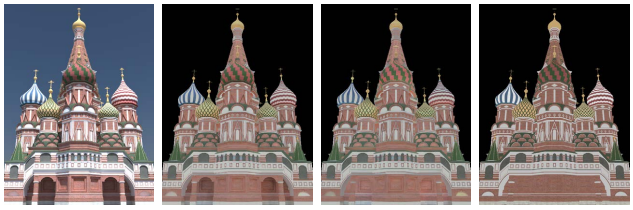
6.2 Analysis and Limitations

Analysis. We show the number of constraints estimated for each scene in Table 1. The size of the downsampled point cloud P_{sel} and the number of candidate pairs for reflectance constraints C_{cand} are approximately the same for all captured and downloaded scenes. However, on average 52% of the pairs are discarded for captured scenes, and 83% for downloaded scenes. Moving from a single-camera, controlled capture setting to online photo collections introduces errors due to different cameras, temporal extent (e.g., repainted façades), and image editing. Our robust statistical criterion detects some of these errors and discards the corresponding pairs.

Fig. 10 shows the effect of correcting pairwise reflectance constraints with the ratios of ambient occlusion (Sec. 4.1). This cor-

	Synth.	Captured		Internet Photo Collections					
Scene	1	2	3	4	5	6	7	8	9
N_{d}	6	5	10	9	11	8	11	8	17
N_{r}	30	32	48	56	60	34	61	28	53
P_{rec}	100k	467k	1.3M	1M	888k	2.0M	1.4M	591k	552k
P_{sel}	68k	200k	199k	200k	196k	192k	199k	200k	196k
C_{cand}	1.4M	1.5M	1.5M	1.3M	1.4M	1.3M	1.5M	1.5M	1.5M
C_{pair}	260k	724k	709k	197k	392k	241k	155k	272k	192k
C_{coher}	39k	105k	142k	66k	65k	57k	46k	54k	53k

Table 1: *N_d Number of images to decompose for each scene, N_r number of images for reflectance ratio estimation (Eq. 3), P_{rec} number of reconstructed 3D points, P_{sel} number of points after downsampling, C_{cand} number of candidate pairs for reflectance constraints before applying statistical criterion, C_{pair} number of reliable pairwise constraints, C_{coher} number of coherency constraints. 1: Synthetic St. Basil; 2: Doll; 3: Temple; 4: St. Basil; 5: Knossos; 6: Moldoviţa; 7: Florence; 8: RizziHaus; 9: Manarola.*

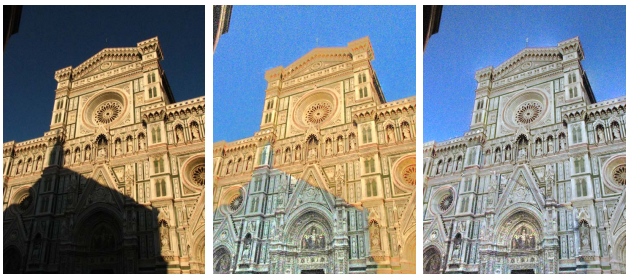


(a) Input rendering (b) Reflectance without correction (c) Reflectance with correction (d) Ground truth reflectance

Figure 10: *Effect of compensating for ambient occlusion on the decomposition of a synthetic image (a). Without special treatment, the reflectance under the arches appears darker (b) because these regions systematically receive less illumination. Correcting the pairwise reflectance constraints by compensating for ambient occlusion (Sec. 4.1) yields a reflectance (c) closer to ground truth (d).*

(a) Illumination without pairwise constraints (b) Illumination with pairwise constraints

Figure 11: *Influence of the pairwise relative constraints on another image of the "Doll scene". (a) Without pairwise reflectance constraints, texture cannot be successfully separated from lighting and the resulting illumination layer contains large texture variations. (b) Enabling these constraints allows recovering a smooth illumination on the tablecloth, despite the complexity of its texture.*



Input image Reflectance without coherence Reflectance with coherence

Figure 12: *Comparison between the decomposition, before and after multi-view coherence in the Florence scene. The coherence constraints between multiple views allow our method to recover a coherent reflectance even under mixed lighting conditions such as this bright sunset with dark blue shadows.*

rection yields a better estimation of reflectance in regions which are systematically in shadow, such as the arches in the synthetic example. Fig. 11 shows the importance of our pairwise constraints for disambiguating reflectance and illumination. In regions with complex texture, they allow us to recover smooth illumination (Fig. 11b), while relying on coherency constraints only results in strong texture artifacts in the illumination (Fig. 11a). In Fig. 12, we first show the decomposition for a single image without the coherence term $E_{\text{coherence}}$, and then the result with coherence constraints to all other images. This image contains challenging mixed lighting conditions, i.e., the blue sky is dominant in the shadow while the bright sun is dominant elsewhere. As a result, the reflectance without coherence constraints contains a residual shadow, which is removed when coherence constraints are added. Additional examples and comparisons can be found in the supplemental materials.

Limitations. We designed our method to estimate coherent reflectance over multiple views of a scene. However, images in photo collections are often captured with different cameras and can be post-processed with different gamma and saturation settings. Since we enforce coherent reflectance, residues of these variations are sometimes visible in our illumination component (e.g., Fig. 8). We argue that *some* reflectance residues in the illumination are acceptable as long as reflectance is plausible and coherent. For example they will be recombined with a coherent (thus similar) reflectance layer when transferring lighting (Sec. 6.3). Correcting for camera responses and image transformations automatically is a promising direction for future work. We expect such corrections to remove the remaining artifacts in our intrinsic image decompositions.

We rely on multi-view stereo for correspondences between views. Consequently in poorly reconstructed regions (such as very dark regions, e.g., just below the roof in Fig. 1), we rely only on the smoothness energy for our decomposition. Since no correspondences exist between views, reflectance in these regions is not coherent across images. If such regions are systematically darker in all views, this is fine for lighting transfer because low illumination values mask the reflectance discrepancy. However, since reflectance is computed by dividing the input image with shading very small shading values can result as very bright pixels in the reflectance. Thin features are also problematic since radiance is blended in the input images. This could be treated with a change of scale, i.e., using close up photos.

6.3 Application to lighting transfer

As an application of our coherent decomposition, we transfer illumination between two pictures of a scene taken from different viewpoints under different illumination (Fig. 13). We use the 3D point cloud as a set of sparse correspondences for which the illumination is known in the two images. We then propagate the illumination of one image to the other image using the smoothness prior of Sec. 5.2. In areas visible only in the target view, the propagation interpolates the illumination values from the surrounding points visible in both images. We generate a radiance image by multiplying the reflectance with the transferred illumination. Since multi-view stereo does not produce 3D points in sky regions, we use the sky detector of Hoiem et al. [2005], correct the segmentation if necessary, and apply standard histogram transfer on sky pixels.

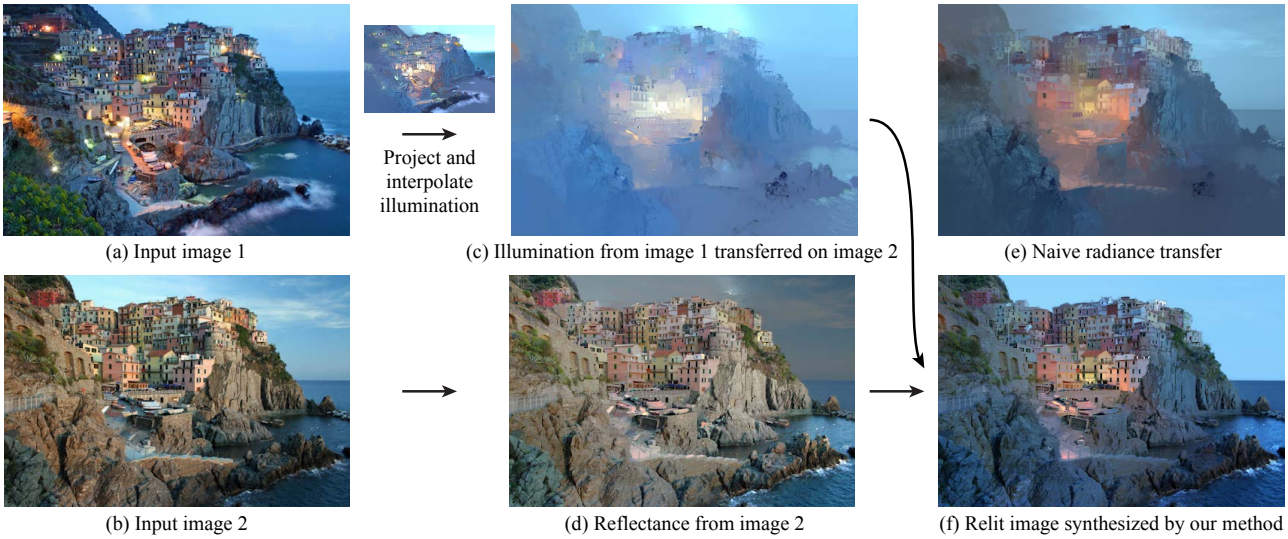


Figure 13: *Given two views of the same scene under different lighting (a,b), we transfer the illumination from one view into the other view (c). We then multiply the transferred illumination by the reflectance layer (d) to synthesize the relit image (f). Transferring the radiance directly fails to preserve the fine details of the reflectance (e).*

In Fig. 13(e) we compare our illumination transfer with direct transfer of radiance. Propagating the radiance produces smooth color variations in-between the correspondences. In contrast, our combination of transferred illumination with the target reflectance preserves fine details. In Fig. 14 we apply our approach to harmonize lighting for multiple viewpoints. In our accompanying video[3] we show image-based view transitions [Roberts 2009] with harmonized photographs. Our method produces stable transitions between views, despite strong shadows in the original images that could not be handled by simple color compensation [Snavely et al. 2008]. We also show artificial timelapse sequences synthesized by transferring all illumination conditions on a single viewpoint.

7 Conclusion

We introduced a method to compute coherent intrinsic image decompositions from photo collections. Such collections contain multiple lighting conditions and can be used to automatically calibrate camera viewpoints and reconstruct 3D point clouds. We leverage this additional information to *automatically* compute coherent intrinsic decompositions over the different views in a collection. We demonstrated how sparse 3D information allows automatic correspondences to be established, and how multiple lighting conditions are effectively used to compute the decomposition. We introduced a complex synthetic benchmark with ground truth, and compared our method to several previous approaches. Our approach outperforms previous methods numerically on the synthetic benchmark and is comparable visually in most cases. In addition, our method ensures that the reflectance layers are coherent among the images. We presented results on a total of 9 scenes and have automatically computed intrinsic image decompositions for a total of 85 images. Our automatic solution shows that the use of coherence constraints can improve the extracted reflectance significantly, and that we can produce coherent reflectance even for images with extremely different lighting conditions, such as night and day. Our coherent intrinsic images enable illumination transfer and stable transitions between views with consistent illumination. This transfer has the potential to benefit to free-viewpoint image-based rendering algorithms that assume coherent lighting when generating novel views from multiple photographs of a scene (e.g., [Chaurasia et al. 2011]).

[3] https://www-sop.inria.fr/reves/Basilic/2012/LBPDD12/

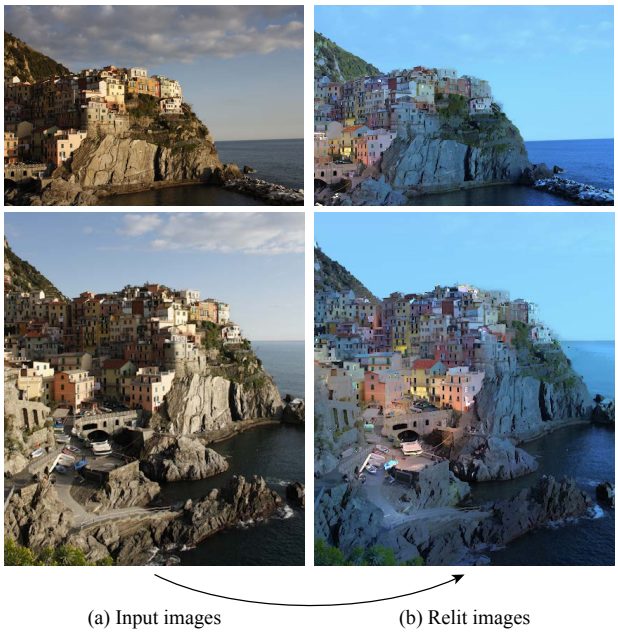


Figure 14: *We use our lighting transfer to harmonize the illumination over multiple images.*

Acknowledgments

This work was partially funded by the EU IP Project VERVE (www.verveconsortium.eu). INRIA acknowledges a generous research donation from Adobe. Fredo Durand acknowledges funding from NSF and Foxconn and a gift from Cognex. We thank

Don Chesnut (Fig. 9), Dawn Kelly (Fig. 12) and the following Flickr users for permission to use their pictures: Fulvia Giannessi and Nancy Stieber (Figs. 1 and 2), Léon Setiani (Fig. 9), SNDahl (Fig. 9), Bryan Chang (Figs. 13 and 14). We thank the authors of other methods for kindly providing comparison images included here. Thanks also to Emmanuelle Chapoulie for modeling support on the synthetic dataset, and Eunsun Lee for help with the capture of indoor scenes.

References

BARROW, H., AND TENENBAUM, J. 1978. Recovering intrinsic scene characteristics from images. *Computer Vision Systems*.

BOUSSEAU, A., PARIS, S., AND DURAND, F. 2009. User-assisted intrinsic images. *ACM Trans. Graph. 28*, 5.

CHAURASIA, G., SORKINE, O., AND DRETTAKIS, G. 2011. Silhouette-aware warping for image-based rendering. *Computer Graphics Forum (Proceedings of the Eurographics Symposium on Rendering) 30*, 4.

DEBEVEC, P., AND ET AL. 2004. Estimating surface reflectance properties of a complex scene under captured natural illumination. Tech. rep., USC Institute for Creative Technologies.

FURUKAWA, Y., AND PONCE, J. 2009. Accurate, dense, and robust multi-view stereopsis. *IEEE Trans. PAMI 32*, 8, 1362–1376.

GARCES, E., MUNOZ, A., LOPEZ-MORENO, J., AND GUTIERREZ, D. 2012. Intrinsic images by clustering. *Computer Graphics Forum*. Eurographics Symposium on rendering, EGSR '12.

GARG, R., DU, H., SEITZ, S. M., AND SNAVELY, N. 2009. The dimensionality of scene appearance. In *IEEE ICCV*, 1917–1924.

GROSSE, R., JOHNSON, M. K., ADELSON, E. H., AND FREEMAN, W. T. 2009. Ground-truth dataset and baseline evaluations for intrinsic image algorithms. In *IEEE ICCV*.

HABER, T., FUCHS, C., BEKAERT, P., SEIDEL, H.-P., GOESELE, M., AND LENSCH, H. 2009. Relighting objects from image collections. In *Proc. IEEE CVPR*, 627–634.

HAYS, J., AND EFROS, A. A. 2007. Scene completion using millions of photographs. *ACM TOG (Proc. SIGGRAPH) 26*, 3.

HOIEM, D., EFROS, A. A., AND HEBERT, M. 2005. Automatic photo pop-up. *ACM TOG (Proc. SIGGRAPH) 24*, 3, 577–584.

HOPPE, H., DEROSE, T., DUCHAMP, T., MCDONALD, J., AND STUETZLE, W. 1992. Surface reconstruction from unorganized points. *SIGGRAPH 26*, 71–78.

HORN, B. K. 1986. *Robot Vision*, 1st ed. McGraw-Hill Higher Education.

LAFFONT, P.-Y., BOUSSEAU, A., AND DRETTAKIS, G. 2012. Rich intrinsic image decomposition of outdoor scenes from multiple views. *IEEE Trans. on Vis. and Comp. Graph.*.

LEE, K. J., ZHAO, Q., TONG, X., GONG, M., IZADI, S., UK LEE, S., TAN, P., AND LIN, S. 2012. Estimation of intrinsic image sequences from image+depth video. In *Proc. ECCV*.

LEVIN, A., LISCHINSKI, D., AND WEISS, Y. 2008. A closed-form solution to natural image matting. *IEEE Trans. PAMI*.

LIU, X., WAN, L., QU, Y., WONG, T.-T., LIN, S., LEUNG, C.-S., AND HENG, P.-A. 2008. Intrinsic colorization. ACM, SIGGRAPH Asia '08, 152:1–152:9.

MATSUSHITA, Y., LIN, S., KANG, S., AND SHUM, H.-Y. 2004. Estimating intrinsic images from image sequences with biased illumination. In *Proc. ECCV*, vol. 3022, 274–286.

MATUSIK, W., LOPER, M., AND PFISTER, H. 2004. Progressively-refined reflectance functions from natural illumination. In *Proc. EGSR*, 299–308.

PHARR, M., AND HUMPHREYS, G. 2010. *Physically Based Rendering: From Theory to Implementation, second edition*. Morgan Kaufmann Publishers Inc.

PREETHAM, A. J., SHIRLEY, P., AND SMITS, B. 1999. A practical analytic model for daylight. In *SIGGRAPH*, 91–100.

ROBERTS, D. A., 2009. Pixelstruct, an opensource tool for visualizing 3d scenes reconstructed from photographs.

SHEN, L., AND YEO, C. 2011. Intrinsic image decomposition using a local and global sparse representation of reflectance. In *Proc. IEEE CVPR*.

SHEN, L., TAN, P., AND LIN, S. 2008. Intrinsic image decomposition with non-local texture cues. In *Proc. IEEE CVPR*.

SHEN, J., YANG, X., JIA, Y., AND LI, X. 2011. Intrinsic images using optimization. In *Proc. IEEE CVPR*.

SNAVELY, N., SEITZ, S. M., AND SZELISKI, R. 2006. Photo tourism: Exploring photo collections in 3d. *ACM TOG (Proc. SIGGRAPH) 25*, 3, 835–846.

SNAVELY, N., GARG, R., SEITZ, S. M., AND SZELISKI, R. 2008. Finding paths through the world's photos. *ACM TOG (Proc. SIGGRAPH) 27*, 3, 11–21.

SUNKAVALLI, K., MATUSIK, W., PFISTER, H., AND RUSINKIEWICZ, S. 2007. Factored time-lapse video. *ACM Transactions on Graphics (Proc. SIGGRAPH) 26*, 3.

TAPPEN, M. F., FREEMAN, W. T., AND ADELSON, E. H. 2005. Recovering intrinsic images from a single image. *IEEE Trans. PAMI 27*, 9.

TROCCOLI, A., AND ALLEN, P. 2008. Building illumination coherent 3d models of large-scale outdoor scenes. *Int. J. Comput. Vision 78*, 2-3, 261–280.

TUITE, K., SNAVELY, N., HSIAO, D.-Y., TABING, N., AND POPOVIC, Z. 2011. Photocity: training experts at large-scale image acquisition through a competitive game. In *Proc. SIGCHI'11*, 1383–1392.

WEISS, Y. 2001. Deriving intrinsic images from image sequences. In *IEEE ICCV*, vol. 2, 68.

WU, C., AGARWAL, S., CURLESS, B., AND SEITZ, S. 2011. Multicore bundle adjustment. In *Proc. IEEE CVPR*, 3057 –3064.

YU, Y., AND MALIK, J. 1998. Recovering photometric properties of architectural scenes from photographs. In *SIGGRAPH'98*.

YU, Y., DEBEVEC, P., MALIK, J., AND HAWKINS, T. 1999. Inverse global illumination: recovering reflectance models of real scenes from photographs. In *SIGGRAPH '99*, 215–224.

ZHAO, Q., TAN, P., DAI, Q., SHEN, L., WU, E., AND LIN, S. 2012. A closed-form solution to retinex with nonlocal texture constraints. *IEEE Trans. PAMI 34*.