An Approximate Shading Model for Object Relighting

Zicheng Liao[⋆†] Kevin Karsch[†] David Forsyth[†]
Zhejiang University[⋆] University of Illinois at Urbana-Champaign[†]

Figure 1. Given an existing 3D scene (left), our method builds approximate object models from image fragments (two teapots and an ostrich) and inserts them in the scene (right). Lighting and shadowing of the inserted objects appear consistent with the rest of the scene. Our method also captures complex reflection (front teapot), refraction (back teapot) and depth-of-field of effects (ostrich). None of these rendering effects can be achieved by image-based editing tools, while our hybrid approach avoids the pain of accurate 3D or material modeling. *3D scene modeling credit to: Jason Clarke*.

Abstract

We propose an approximate shading model for image-based object modeling and insertion. Our approach is a hybrid of 3D rendering and image-based composition. It avoids the difficulties of physically accurate shape estimation from a single image, and allows for more flexible image composition than pure image-based methods. The model decomposes the shading field into (a) a rough shape term that can be reshaded, (b) a parametric shading detail that encodes missing features from the first term, and (c) a geometric detail term that captures fine-scale material properties. With this object model, we build an object relighting system that allows an artist to select an object from an image and insert it into a 3D scene. Through simple interactions, the system can adjust illumination on the inserted object so that it appears more naturally in the scene. Our quantitative evaluation and extensive user study suggest our method is a promising alternative to existing methods of object insertion.

1. Introduction

An important task of image composition is to take an existing *image fragment* and insert it into another scene. This approach is appealing because 3D models are difficult to build, and image fragments carry real texture and material effects that achieve realism in a data-driven manner (Fig. 1).

Relighting is generally necessary in the process. To relight an object from an image fragment, we need to know its shape and material properties. Alternatively, Image-based composition methods [6, 24, 1] skips the relighting process, relying on the artist's discretion for determining shading-compatible image fragments. This limits the range of data that can be used for a particular scene. Despite the limitations, image-based methods have been widely used, because shape estimation (required in the other approach) remains a very challenging task. State-of-the-art algorithms, such as the SIRFS method of Barron et al. [2], still produce weak shapes and do not work well for complex materials on real world data. Is there a compromise between the two spaces?

We propose such an approach by exploring an *approximate shading model*. The model circumvents the formidable 3D reconstruction problem, yet is reflexible and can adapt to various target scenes lighting condition.

The model is inspired by two lines of work. First, it is not currently possible to produce veridical reconstructions from single images; but studies show the human visual system is tolerant of renderings that are not physical ([23, 26, 9, 7] give some rules that must apply to an image for it to be seen as accurate). Our model attempts to exploit this fact to conceal weak reconstructions, following the spirit of work in material editing [18] and illumination estimation [16]. Second, illumination cone theory [5] suggests an accurate

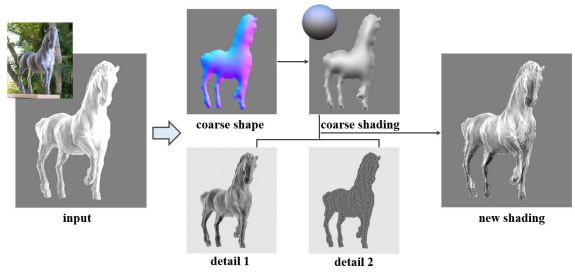


Figure 2. Given an image fragment with albedo-shading decomposition (left, albedo is omitted in this illustration), we build a coarse shape from contour; we then derive two shading detail layers (*detail 1* and *detail 2*) from the shading image and the coarse shape (middle). Our approximate shading model reshades the object under new illumination and produce a new shading on the right. Notice the change in both gross shading (new lighting from above) and the surface detail. The detail images are rescaled for visualization.

shading can be expressed as a linear combination of a few components. Our method decomposes shading into 3 components – a smooth component captured by a coarse shape h, and two shading detail layers: parametric residual S_p and geometric detail S_g. A new shading is expressed as the coarse shading plus a weighted combination of the two detail layers:

$$S(h, S_p, S_g) = shade(h, L), + w_p S_p + w_g S_g \qquad (1)$$

where L is (new) illumination, w_p and w_g are scalar weights (Fig. 2). We refer to S_p as *detail 1* and S_g as *detail 2* throughout the paper.

The coarse shape produces a smoothing shading that captures directional and coarse-scale illumination effects that are critical for perceptual consistency. The shape is purely based on contour constraint (shape from contour), easy to construct and robust to generic views. The two detail layers (S_p and S_g) account for visual complexity of the object. They encode mid and high frequencies of the shading signal left out by the smooth shading component respectively. Intuitively, the mid frequency shading corresponds to object-level features (silhouettes, creases and folds, etc.), and the high frequency shading to material properties. Notice the image-based composition of the detail layers is not physically-based. In practice, however, it yields remarkably good results.

We conduct extensive evaluations of the model by quantitative measurement and user study of visual realism. All results indicate that our model is a promising alternative to existing methods for object insertion. The quantitative measurement is a re-rendering MSE on the augmented MIT intrinsic image dataset, compared with state-of-the-art shape reconstruction method by Barron and Malik [2]. Our model yielded slightly lower MSE on the Lab illumination set. The user study is more compelling as visual realism is the primary goal of image composition. The study showed that subjects preferred our results over that of Barron and Malik by a margin of 20%, and over that of Karsch et al. [16] (synthetic models) by a margin of 14%. Another user study showed the effectiveness of the two detail layers: as more detail layers were applied, the results were preferred by a significantly larger (20-30%) percentage of human subjects.

2. Related work

Object insertion takes an object from one source and sticks it into a target scene. Pure image-based methods [6, 24, 1] totally rely on artist's discretion for shading consistency. [19, 8] take a data driven approach, searching in a large database for compatible source. A relighting procedure would expand the range of images to composite with, because shading inconsistency can be taken care of. Khan et al. [18] show a straightforward method that simulate changes in the apparent material of objects, given an approximate normal field and environment map. Recently, Karsch et al. [16, 17] introduce a technique that reconstructs a 3D scene from a single indoor image and allows synthetic objects to be inserted in.

Our object insertion system takes Karsch et al.'s technique [16] for scene modeling. The major advance is that the inserted objects are taken from images instead of existing 3D models. This expands the source of objects to use in application. Besides, our image-based approach inherits the intrinsic power of the data-driven method, allowing complex material and detail to be easily preserved; with synthetic model these effects are compromised.

Shape estimation Current methods are still unable to recover accurate shape from a single image, even inferring satisfactory approximate shape is difficult. Methods that recover *shape from shading* (SfS) are unstable as well as inaccurate, particularly in the absence of reliable albedo and illumination [29, 11, 12]. A more sophisticated approach is to jointly estimate shape, illumination and albedo from a single image [2]. An alternative is to recover *shape from contour* cues alone, for example, by building a surface that is smooth and compelled to have normal constraints along the contour [15, 25]. Yet another alternative is to assume that lighter pixels are nearer and darker pixels are further away [18].

For re-rendering and other predictive purposes, an alternative to a shape estimate would be **illumination cone** (a representation of all images of an object in a fixed configuration, as lighting changes). This cone is known to lie close to a low dimensional space [4] (a 9-Spherical Harmonics illumination can account for up to 98% of shading variation), suggesting that quite low-dimensional image based reshading methods are available. Our representation could be seen as a hybrid of a shape and an illumination cone representation.

Material and Illumination Our method needs to decompose an input image into albedo and shading (*intrinsic images* [3]). The standard Retinex method [20] assumes sharp changes come from albedo, and slow changes come from shading. We use a variant of the color Retinex algorithm [13] for albedo and shading estimation. Liao et al. [21] propose a method that decomposes an image into albedo, smooth shading, and shading detail caused by high frequency geometry. We use this method to derive our geometric detail layer. We assume illumination is part of a 3D scene, or automatically recovered from image [22, 16].

3. Creating the model

Our object model has four components. We compute a coarse 3D shape estimate, then compute three maps: the albedo, a parametric shading residual, and a geometric detail layer. We refer to the "coarse shading" by the shape, the "parametric shading residual" and the "geometric detail" as the three shading components.

3.1. Coarse shape

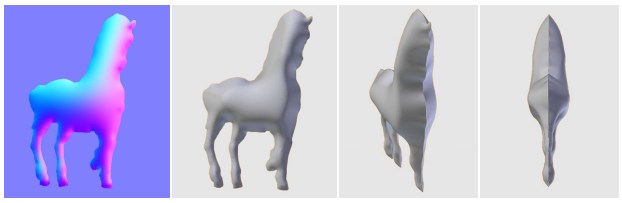
We assume the object to be inserted is an image fragment, and wish to estimate what its appearance is under new illumination. Exact shape is ideal but unavailable. We need a representation capable of capturing gross shading effects. For example, a vertical cylinder with light from the left will be light on left, dark on right. Moving the light to the right will cause it to become dark on left, light on right. We also want our reconstruction to be consistent with a generic view assumption. This implies that (a) the outline should not shift too much if the view shifts, and (b) there should not be large bumps in the shape that are concealed by the view direction (Fig. 11 demonstrates these kinds of mistakes typically generated by SfS methods). To support these, we use a simple shape from contour (SFC) method with stable outline and smooth surface (Fig. 3).

First, we create a normal field by constraining normals on the object boundary to be perpendicular to the view direction, and interpolate them from the boundary to the interior region, similar to Johnston's Lumo technique [15]. Let N be the normal field, Ω and $\partial\Omega$ be the set of pixels in the object and on boundary, respectively. We compute N by the following optimization:

$$\begin{aligned} \min_{N} \quad & \sum_{\Omega} ||\nabla \mathbf{N}||^2 + (||\mathbf{N}|| - 1)^2 \\ \text{subject to} \quad & N_z^i = 0 \quad \forall i \in \partial\Omega \end{aligned} \tag{2}$$

We then reconstruct an approximate height field h from the normal by minimizing:

$$\sum_{\Omega} ||(\frac{\partial h}{\partial x} - \frac{N_x}{max(\epsilon, N_z)})||^2 + ||(\frac{\partial h}{\partial y} - \frac{N_y}{max(\epsilon, N_z)})||^2 \tag{3}$$



(a) normal (b) 3D shape (c) view2 (d) view3

Figure 3. Shape reconstruction. Our reconstruction is simple but robust to large errors that may occur in state-of-the-art SfS algorithm and supports the generic view assumption (Fig. 11).

subject to $h_i = 0$ for boundary pixels (stable outline). The threshold avoids numerical issues near the boundary for exact reconstruction (Wu et al. [27]); and forces the reconstructed object to have a crease along its boundary. This crease is very useful for the support of generic view direction, as it allows slight change of view direction without exposing the back of the object and causing self-occlusion. The reconstructed height field is then flipped to make a symmetric full 3D shape (Fig. 3).

3.2. Parametric shading residual

The coarse shape can recover gross changes in shading caused by lighting. However, it cannot represent finer detail. We use shading detail maps to represent this detail. We define the shading detail maps as a representation of the residual incurred by fitting the object shading estimate with some model. We use two shading details in our model: *parametric shading residual* that encodes object level features (silhouettes, crease and folds, etc.), and *geometric detail* that encodes fine scale material effects.

First, we use a standard color Retinex algorithm [13] to get an initial albedo ρ and shading S estimate from the input image: $I = \rho \cdot S$. We then use a parametric illumination model $L(\theta)$ to shade the coarse shape and compute the residual by solving:

$$\hat{\theta} : \underset{\theta}{\text{argmin}} \sum ||S - \mathit{shade}(h, L(\theta))||^2 \tag{4}$$

The optimized illumination $\hat{\theta}$ is substituted to obtain the parametric shading residual:

$$S_p = S - \text{Shade}(h, L(\hat{\theta})). \tag{5}$$

Many parametric illuminations are possible (i.e., spherical harmonics). We used a mixture of 5 point sources, the parameters being the position and intensity of each source, forming a 20-dimensional representation.

Figure 4 upper right row shows an example of the best fit coarse shading and the resultant parametric shading detail. Note that the directional shading is effectively removed, leaving shading cues of object level features. The bottom right row shows the geometric detail extraction pipeline.

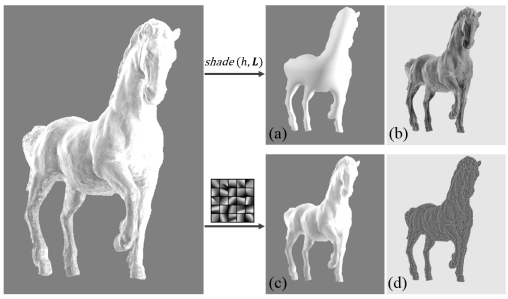


Figure 4. Illustration of shading detail computation. Given the shading image (left), the upper right row shows the parametric fitting procedure that computes the best fit shading (a) from the shape and the parametric shading residual (b); the bottom right row shows the non-parametric patch-based filtering procedure that computes the filtered shading image (c) and the residual known as geometric detail (d).

3.3. Geometric detail

The parametric shading residual is computed by a *global* shape and illumination parameterization, and contains all the shading details missed by the shape. Now we wish to compute another layer that contains only fine-scale details. We use a technique from Liao et al. [21], in which they extract very fine-scale geometric details with a *local* patch-based non-parametric filter. The resultant geometric detail represents high frequency shading signal caused by local surface geometry like bumps and grooves and is insensitive to gross shading and higher-level object features such as silhouettes. See the difference to the parametric shading residual in Fig. 4.

The filtering procedure uses a set of shading patches learned from smooth shading images to reconstruct an input shading image. Because geometric detail signals are poorly encoded by the smooth shading dictionary, they are effectively left out. In the experiment we use a dictionary of 500 patches with patch size 12×12.

How many detail layers are necessary? We choose two layers empirically as this compromises representational power for ease of editing. On the one hand, it is entirely reasonable to increase the number of detail layers for representational power. However, this would make the editing interface less easy to use (section 4, Fig. 5). On the other hand, having two detail layers allows user to adjust mid (eg. a muscle) and high (eg. a wrinkle) frequency shadings separately, simulating physical shading changes in a more flexible way.

4. A relighting system

With the object model, we develop a system that relights an object from image into a new scene. The system combines interactive scene modeling, physically-based rendering and image-based detail composition (Fig. 5).

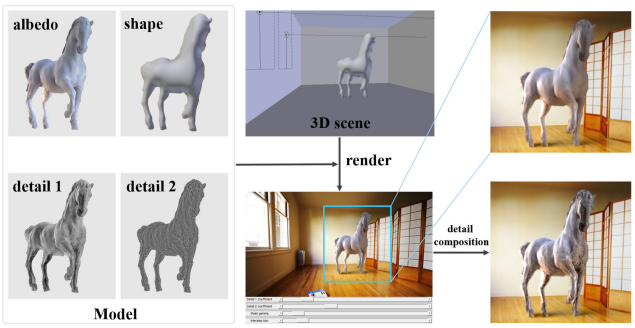


Figure 5. Illustration of the relighting system. Given an object model (the horse), an artist places the model into a 3D scene, render it with a physically-based renderer, and then composite it with the detail layers to generate the final result (a close-up view in bottom right corner). Notice the difference of surface detail appearance on the horse before and after detail composition.

4.1. Modeling and Rendering

We build a sparse mesh object from the height field (by pixel-grid triangulation and mesh simplification [28]) for the object. The target scene can be existing 3D scenes, or built from an image (Karsch et al. [16], Hedau et al. [14], etc.). The artist then selects an object and places it into the scene, adjusting its scale and orientation, and making sure the view is roughly the same as that of the object in the original image. The model is then rendered with the estimated albedo. For all the results in the paper, we use Blender (http://blender.org) for modeling and LuxRender (http://luxrender.net) for physically-based rendering. All target scenes were constructed using the technique of Karsch et al. [16] technique if not otherwise stated.

To create the mesh object for rendering, We flip the mesh along the contour plane to create a closed 3D mesh (Fig. 3). However, the flipped shape model is thin along the base and can cause light leaks and/or skinny lateral shadows. We use a simple user-controllable extrusion procedure to handle such cases. Besides, our shape model assumes an orthographic camera, while most rendering systems use a perspective camera. This will cause texture distortion during texture mapping. We use a simple "easing" method to avoid it. Since rendering is not the emphasis of this work, we refer interested readers to supplemental material "Easing and back extrusion" section for more details.

4.2. Detail composition

We then composite the rendered scene with the two detail maps and original scene to produce final result (Fig. 6).

First, we composite the two shading detail images with the shading field of the rendered image:

$$C = \rho(S + w_p S_p + w_g S_g) \tag{6}$$

Figure 6. Object relighting by our method (right column), comparing with the method using shape estimates by Barron and Malik [2] (middle column, missing shadows due to its "flat" shape reconstruction). Our relighting method adjusts the shading on the object for a variety of scenes with different illumination conditions. Detail composition simulates complex surface geometry and materials properties that is difficult to achieve by physically-based modeling. Best viewed in color at high-resolution.

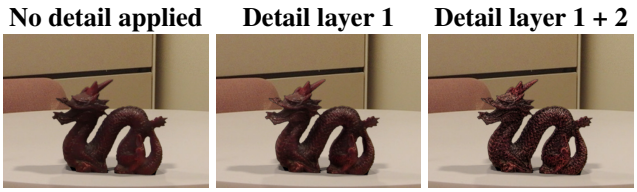


Figure 7. Detail composition. The left column displays relighting results with our coarse shape model and estimated albedo. The middle column displays results compositing with only the parametric shading residual. Notice how this component adds object level shading cues and improves visual realism. The right column are results compositing with both detail layers. Fine-scale surface detail is further enhanced (see the dragon). Best viewed in color at high-resolution.

where $S = I_r/\rho$ (equivalently denoted as $shade(h, L)$ in equation 1) is the shading field, I_r is the rendered image. The weights w_p and w_g can be automatically determined by regression (section 5.1) or manually adjusted by artist with a slider control (Fig. 5 detail composition).

Second, we use standard techniques (e.g. [10, 16]) to composite C with the original image of the target scene. This produces the final result. Write I_t for the target image, I_e for the empty rendered scene *without* the inserted object, and M for the object matte (0 where no object is present, and $(0, 1]$ otherwise). The final composite image C is obtained by:

$$C_{\text{final}} = M \odot C + (1 - M) \odot (I_t + I_r - I_e). \quad (7)$$

Compositing the two detail layers improves the visual realism of the rendered object (Fig. 7). A controlled user study (section 5.2, task 1) showed that users consistently prefer composition results with more detail layers applied.

5. Evaluation

Our assumption is that the approximate shading model can capture major effects of illumination change of an object in new environment and generate visually plausible image. To evaluate the performance, we compare our representation with state-of-the-art shape reconstructions by Barron and Malik [2] on a re-rendering metric (Sec. 5.1). We also conduct an extensive set of user study to evaluate the realism of our relighting results versus that of Barron and Malik [2], Karsch et al. [16] and real scenes (Sec. 5.2). The evaluation results show that our representation is a very promising alternative to the existing methods for object relighting.

5.1. Re-rendering Error

The re-rendering metric measures the error of relighting an estimated shape. On a canonical shape representation (a depth field), the metric is defined as

$$\text{IMSE}_{rerender} = \frac{1}{n}||I - k\hat{\rho}\,\text{ReShade}(\hat{h}, L)||^2 \quad (8)$$

where $\hat{\rho}$ and $\hat{h}$ are estimated albedo and depth, I is the re-rendering with the ground truth shape h^* and albedo ρ^*: $I = \rho^*\text{ReShade}(h^*, L)$, n is the number of pixels, k is a scaling factor that minimizes the squared error.

With our model, write $S_c = shade(h, L), S_p, S_g$ for the coarse shading, parametric shading detail and the geometric detail, respectively, and replace the corresponding part of Equation 1 with $\text{ReShade}(S(L), w) = S_c + w_p S_p + w_g S_g$ for some choice of weight vector $w = (1, w_p, w_g)$. The re-rendering metric is:

$$\text{IMSE}'_{rerender} = \frac{1}{n}||I - k\hat{\rho}\,\text{ReShade}(S(L), w)||^2 \quad (9)$$

That is, rendering of canonical shape is replaced by our approximate shading model (Equation 1).

We offer three methods to select w. An *oracle* could determine the values by least square fitting that leads to best MSE. *Regression* could offer a value based on past experience. We learn a simple linear regression model to predict the weights from illumination. Lastly, an artist could *manually* choose the weights, as demonstrated in our relighting system (Sec. 4.2).

Experiment We run the evaluation on the augmented MIT Intrinsic image dataset [2]. To generate the target images, we re-render each of the 20 object by 20 randomized monochrome (9×1) Spherical Harmonics illuminations, forming a 20×20 image set. We then measure the re-rendering error of our model and Barron and Malik's reconstructions. For our method, we compare models built from the Natural Illumination dataset and Lab Illumination dataset separately. The models are built (a) in the default setting, (b) using Barron and Malik's shading estimation, and (c) using the ground truth shading. See Table 1 for the results. To learn the linear regression model, we draw for each object 100 nearest neighbors in illumination space from the other 380 data points, and fit its weights by LSQ.

Method	"Natural" Illum. (No strong shadows)		Lab Illum. (With strong shadows)	
B & M [2]	**0.017**		0.037	
Ours	*LSQ*	*Reg.*	*LSQ*	*Reg.*
(a) default	0.033	0.036	0.059	0.064
(b) $S1$	0.027	0.032	**0.034**	**0.036**
(c) $S2$	0.021	0.024	0.023	0.024

Table 1. Re-rendering error of our method compared to Barron & Malik [2]. Our method performs less as good on a synthetic image dataset (the "Natural" Illumination dataset) where Barron and Malik produces very accurate shape estimates. On the real image dataset (the "Lab" Illumination dataset), our method with automatic weights can perform better than Barron and Malik.

Table 1 displays our experiment result. The result shows that when the shape estimation of Barron and Malik is accurate (on the "Natural" Illumination dataset, a synthetic dataset by the same shading model used in their optimization), our approximate shading performs less well. This is reasonable because a perfect shape is supposed to produce zero error in the re-rendering metric. This is also acceptable because the dataset images are not real. On the real image set (the "Lab" Illumination dataset, taken in lab environment with strong shadows), the shape estimation of Barron and Malik becomes inaccurate, and our approximate shading model can produce lower error with both regressed weights and oracle setting. With better detail layers (when ground truth shading is used to derive them), our model achieves significantly lower errors.

It is worth noting that MSE is not geared toward *visual realism* of insertion results (image features takes little weight; non-linearity of visual perception on light intensity, etc.). To evaluate that, we further conduct a set of user studies as follows.

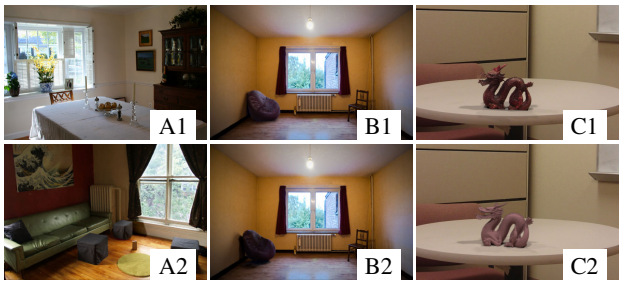


Figure 8. Example trial pairs from our user study. The left column shows an insertion result by our method and a real image (task 1). The middle column shows that from our method and the method of Barron and Malik (task 3). And the right column shows that from our method and the method of Karsch et al. [16]. Users were instructed to choose the picture from the pair that looked the most realistic. For each row, which image would you choose? Best viewed in color at high-resolution.

Our results: A2, B1, C1 (ottoman inserted near window in A2);

5.2. User study

In the study, each subject is shown series of two-alternative forced choice tests and chooses between each pair which he/she feels the most realistic. We tested five different tasks: (1) our method with controlled number of detail layers, (2) our method against real images, (3) the method of Barron and Malik against real images, (4) our method against Barron and Malik [2], and (5) our method against Karsch et al. [16]. The first task shows both detail layers help to make more visually realistic result. The other four tasks reveal the advantage of our result over that of Barron and Malik [2] and Karsch et al. [16]. Figure 8 shows example trials from the first, third and fourth tasks.

Experiment setup For each task, we created 10 different insertion results using a particular method (ours, Barron and Malik, or Karsch et al. For the results of Barron and Malik, we ensured the same object was inserted at roughly the same location as our results. This was not the case for the results of Karsch et al., as synthetic models were not all available for the objects we chose. We also collected 10 real scenes (similar to the ones with insertion) for the tasks involving real images. Each subject viewed all 10 pairs of images for one but only one of the five tasks. For the 10 results by our method, the detail layer weights were manually selected (it is hard to apply the regression model as in Section 5.1 to the real scene illuminations) while the other two methods do not have such options.

We polled 100-200 subjects using Mechanical Turk for each task. In an attempt to avoid inattentive subjects, each task also included four "qualification" image pairs (a cartoon picture next to a real image). Subjects who incorrectly chose any of the four cartoon picture as realistic were removed from our findings (6 in total, leaving 294 studies with usable data). At the end of the study, we showed subjects two additional image pairs: a pair containing rendered spheres (one a physically plausible, the other not), and a pair containing line drawings of a scene (one with proper vanishing point perspective, the other not). For each pair, subjects chose the image they felt looked most realistic. Then, each subject completed a brief questionnaire, listing demographics, expertise, and voluntary comments.

These answers allowed us to separate subjects into subpopulations: **male/female**, **age** < 25 **/** ≥ 25, whether or not the subject correctly identified both the physically accurate sphere *and* the proper-perspective line drawing at the end of the study (**passed/failed perspective-shading (p-s) tests**), and also **expert/non-expert** (subjects were classified as experts only if they passed the perspective-shading tests *and* indicated that they had expertise in art/graphics). We also attempted to quantify any learning effects by grouping responses into the **first half** (first five images shown to a subject) and the **second half** (last five images shown).

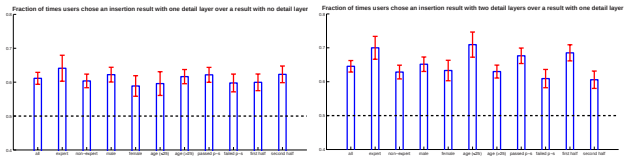


Figure 9. User rating on our results with controlled number of detail layers. Users consistently prefer insertion results with more detail layers. In the first figure, users prefer results with the parametric detail over that with no detail layer applied in 61% of 764 viewed pairs. In the second one, users prefer results with both parametric detail and geometric detail over that with only the parametric detail in 65% of 756 viewed pairs.

Results and discussion In task 1, users consistently preferred insertion results with more detail layers applied (Figure 9). In the other four tasks, the user study showed that subjects confused our insertion result with a real image in 44% of 1040 viewed pairs (task 2, see table 2); an optimal result would be 50%. We also achieve better confusion rates than the insertion results of Barron and Malik [2] (task 3, 42%), and perform well ahead of the method of Barron and Malik in a head-to-head comparison (task 4, Fig. 10 left), as well as a head-to-head comparison with the method of Karsch et al. [16] (task 5, Fig. 10 right).

Figure 9 displays the result of task 1. It demonstrates the model makes better insertion results as more detail layers are applied. Overall, users preferred insertion results with detail 1 over that without detail composition in 61% of 764 viewed image pairs, and preferred results with both detail layers over that with only detail 1 in 65% of 756 viewed pairs. Consistent results were shown in all subpopulations.

Table 2 demonstrates how well images containing inserted objects (using either our method or Barron and Malik) hold up to real images (tasks 2 and 3). We observe better confusion rates (e.g. our method is confused with real images more than the method of Barron and Malik) overall and in each subpopulation except for the population who failed the perspective and shading tests in the questionnaire.

We also compared our method and the method of Barron and Malik head-to-head by asking subjects to choose a more realistic image when shown two similar results side-by-side (Fig. 8 middle column). Figure 10 summarizes our findings. Overall, users chose our method as more realistic in a side-by-side comparison on average 60% of the time in 1000 trials. In all subject subpopulations, our method was preferred by a large margin to the method of Barron and Malik; each subpopulation was at least two standard deviations away from being "at chance" (50% – see the red bars and black dotted line in Fig. 10). Most interestingly, the expert subpopulation preferred our method by an even greater margin (66%), indicating that our method may appear more realistic to those who are good judges of realism.

Karsch et al. [16] performed a similar study to evaluate their 3D synthetic object insertion technique, in which subjects were shown similar pairs of images, except the inserted objects were synthetic models. In their study, subjects chose the insertion results only 34% of the time, much lower than the two insertion methods in this study, a full 10 points lower than our method and 8 points lower than the method of Barron and Malik. While the two studies were not conducted in identical setting, the results are nonetheless intriguing. We postulate this large difference is due to the nature of the objects being inserted: we use *real* image fragments that were formed under real geometry, complex material and lighting, sensor noise, and so on; they use 3D models for which photorealism can be extremely difficult to model. By inserting image fragments instead of 3D models, we gain photorealism in a data-driven manner (Fig. 8 C1 versus C2). This postulation is validated by our comparison in task 5. For all but one subpopulations, our results were preferred by a large margin (Fig. 10 right).

Subpopulation	#trials	ours (%)	B & M [2] (%)
all	1040	**44.0±1.5**	41.8±1.6
expert	200	**43.5±3.4**	36.2±3.7
non-expert	840	**44.2±1.7**	42.9±1.8
passed p-s test	740	**44.2±1.8**	40.5±2.1
failed p-s test	300	43.7±2.7	**43.9±2.5**
male	680	**44.7±1.9**	42.6±1.8
female	360	**42.8±2.4**	39.0±3.5
age $\leq$ 25	380	**43.2±2.5**	41.7±2.5
age >25	660	**44.5±1.9**	41.9±2.0
1st half	520	**45.6±2.2**	43.0±2.3
2nd half	520	**42.5±2.0**	41.8±2.1

Table 2. Fraction of times subjects chose an insertion result over a real image in the study. Overall, users confused our insertion results with real pictures 44% of the time, while confusing the results of Barron and Malik with real images 42% of the time. Interestingly, for the subpopulation of "expert" subjects, this difference became more pronounced (44% vs 36%). Each cell shows the mean standard deviation.

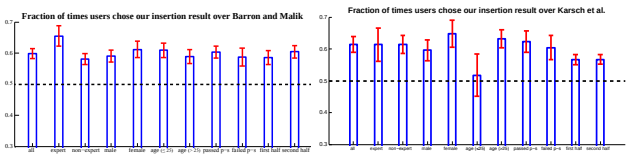


Figure 10. Left: Comparison of our results against that by the method of Barron and Malik. our results were chosen as more realistic in 60% of the trials ($N = 1000$). For all subpopulations, our results were preferred well ahead of the other as well. All differences to the dotted line (equal preference) are greater than two standard deviation. The "expert" subpopulation chose our insertion results most consistently. Right: Comparison against the method of Karsch et al. The advantage our method holds over that of Karsch et al. is similar to the advantage over Barron and Malik. Number of trials = 1840.

(a) Input (b) Shape (c) B & M (d) Our result

Figure 11. In this example, we built models from the cube in the input image (cyan box) and inserted it back into the scene. Sophisticated SfS methods (in this case, Barron and Malik [2]) can have large error and unstable boundaries that violates the generic view assumption. For object insertion, lighting *on* the object is important, but it is equally important that cast shadows and interreflected light look correct; shape errors made by complex SfS methods typically exacerbate errors both *on* and *around* the object (see cast shadows in c). Our shape is simple but behaves well in many situations and is typically robust to such errors (d). Best viewed in color at high-resolution.

Acknowledgements This work was supported in part by NSF under Grants No. NSF IIS 09-16014, IIS-1421521, and ONR MURI Award N00014-10-10934; and in part by 973 Grant No. 2015CB352302 and ZJNSF Grant No. Q15F020006.

References

[1] A. Agarwala, M. Dontcheva, M. Agrawala, S. Drucker, A. Colburn, B. Curless, D. Salesin, and M. Cohen. Interactive digital photomontage. *ACM Trans. Graph.*, 23(3):294–302, 2004. 1, 2

[2] J. T. Barron and J. Malik. Color constancy, intrinsic images, and shape estimation. *ECCV*, 2012. 1, 2, 5, 6, 7, 8

[3] H. Barrow and J. Tenenbaum. Recovering intrinsic scene characteristics from images. In *Comp. Vision Sys.*, pages 3–26, 1978. 3

[4] R. Basri and D. Jacobs. Lambertian reflectance and linear subspaces. *PAMI*, 2003. 2

[5] P. N. Belhumeur and D. J. Kriegman. What is the set of images of an object under all possible illumination conditions? *IJCV*, 1998. 1

[6] P. J. Burt and E. H. Adelson. The laplacian pyramid as a compact image code. *Communications, IEEE Transactions on*, 31(4):532–540, 1983. 1, 2

[7] P. Cavanagh. The artist as neuroscientist. *Nature*, pages 301–307, 2005. 1

[8] T. Chen, M.-M. Cheng, P. Tan, A. Shamir, and S.-M. Hu. Sketch2photo: internet image montage. *ACM Trans. Graph.*, 28(5):124:1–124:10, Dec. 2009. 2

[9] B. Conway and M. Livingstone. Perspectives on science and art. *Current Opinion in Neurobiology*, 17:476–482, 2007. 1

[10] P. Debevec. Rendering synthetic objects into real scenes: bridging traditional and image-based graphics with global illumination and high dynamic range photography. In *SIGGRAPH'98*, pages 189–198. ACM, 1998. 5

[11] J.-D. Durou, M. Falcone, and M. Sagona. Numerical methods for shape-from-shading: A new survey with benchmarks. *Comput. Vis. Image Underst.*, 109(1):22–43, 2008. 2

[12] D. Forsyth. Variable-source shading analysis. *IJCV'11*, 91, 2011. 2

[13] R. Grosse, M. K. Johnson, E. H. Adelson, and W. T. Freeman. Ground-truth dataset and baseline evaluations for intrinsic image algorithms. In *ICCV*, pages 2335–2342, 2009. 3

[14] V. Hedau, D. Hoiem, and D. A. Forsyth. Recovering the spatial layout of cluttered rooms. In *ICCV*, pages 1849–1856. IEEE, 2009. 4

[15] S. F. Johnston. Lumo: illumination for cel animation. In *NPAR '02*, 2002. 2, 3

[16] K. Karsch, V. Hedau, D. Forsyth, and D. Hoiem. Rendering synthetic objects into legacy photographs. In *ACM Trans. Graph (SIGGRAPH Asia)*, volume 30, pages 157:1–157:12, 2011. 1, 2, 3, 4, 5, 6, 7

[17] K. Karsch, K. Sunkavalli, S. Hadap, N. Carr, H. Jin, R. Fonte, M. Sittig, and D. Forsyth. Automatic scene inference for 3d object compositing. *ACM Trans. Graph.*, 33(3):32:1–32:15, June 2014. 2

[18] E. A. Khan, E. Reinhard, R. W. Fleming, and H. H. Bülthoff. Image-based material editing. *ACM Trans. Graph.*, 25(3):654–663, July 2006. 1, 2

[19] J.-F. Lalonde, D. Hoiem, A. A. Efros, C. Rother, J. Winn, and A. Criminisi. Photo clip art. *ACM Transactions on Graphics (SIGGRAPH 2007)*, 26(3):3, August 2007. 2

[20] E. Land and J. McCann. Lightness and retinex theory. In *JOSA*, volume 61, pages 1–11, 1971. 3

[21] Z. Liao, J. Rock, Y. Wang, and D. Forsyth. Non-parametric filtering for geometric detail extraction and material representation. In *CVPR*, pages 963–970, Washington, DC, USA, 2013. IEEE Computer Society. 3, 4

[22] J. Lopez-Moreno, S. Hadap, E. Reinhard, and D. Gutierrez. Compositing images through light source detection. *Computers Graphics*, 34(6):698 – 707, 2010. ¡ce:title¿Graphics for Serious Games¡/ce:title¿ ¡ce:title¿Computer Graphics in Spain: a Selection of Papers from {CEIG} 2009¡/ce:title¿ ¡ce:title¿Selected Papers from the {SIGGRAPH} Asia Education Program¡/ce:title¿. 3

[23] Y. Ostrovsky, P. Cavanagh, and P. Sinha. Perceiving illumination inconsistencies in scenes. *Perception*, 34:1301–1314, 2005. 1

[24] P. Pérez, M. Gangnet, and A. Blake. Poisson image editing. *ACM Trans. Graph.*, 22(3):313–318, July 2003. 1, 2

[25] M. Prasad and A. Fitzgibbon. Single view reconstruction of curved surfaces. In *CVPR*, pages 1345–1354, 2006. 2

[26] P. Sinha, B. Balas, Y. Ostrovsky, and R. Russell. Face recognition by humans: Nineteen results all computer vision researchers should know about. *IEEE*, 94(11):1948–1962, 2006. 1

[27] T.-P. Wu, J. Sun, C.-K. Tang, and H.-Y. Shum. Interactive normal reconstruction from a single image. *ACM Trans. Graph.*, 27(5):119:1–119:9, Dec. 2008. 3

[28] T. Xia, B. Liao, and Y. Yu. Patch-based image vectorization with automatic curvilinear feature alignment. *ACM Trans. Graph.*, 28(5):115:1–115:10, Dec. 2009. 4

[29] R. Zhang, P.-S. Tsai, J. Cryer, and M. Shah. Shape-from-shading: a survey. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 21(8):690–706, 1999. 2