Leveraging Multi-View Image Sets for Unsupervised Intrinsic Image Decomposition and Highlight Separation

Renjiao Yi,[1] **Ping Tan,**[1] **Stephen Lin**[2]
[1]Simon Fraser University, Burnaby, Canada
[2]Microsoft Research, Beijing, China
{renjiaoy, pingtan}@sfu.ca, stevelin@microsoft.com

Abstract

We present an unsupervised approach for factorizing object appearance into highlight, shading, and albedo layers, trained by multi-view real images. To do so, we construct a multi-view dataset by collecting numerous customer product photos online, which exhibit large illumination variations that make them suitable for training of reflectance separation and can facilitate object-level decomposition. The main contribution of our approach is a proposed image representation based on local color distributions that allows training to be insensitive to the local misalignments of multi-view images. In addition, we present a new guidance cue for unsupervised training that exploits synergy between highlight separation and intrinsic image decomposition. Over a broad range of objects, our technique is shown to yield state-of-the-art results for both of these tasks.

Introduction

Separating reflectance layers in an image is an essential step for various image editing and scene understanding tasks. One such layer is composed of highlights, which are mirror-like reflections off the surface of objects. Extracting highlights from an image can be useful for problems such as estimating scene illumination (Lombardi and Nishino 2016; Yi et al. 2018) and reducing the oily appearance of faces (Li, Zhou, and Lin 2015). The other two layers represent shading and albedo. Their separation is commonly known as intrinsic image decomposition, which has been utilized in applications such as shading-based scene reconstruction (Yu et al. 2013; Or-El et al. 2015) and texture replacement in images (Weiss 2001; Jeon et al. 2014).

Factorizing an image into the three reflectance layers is an ill-posed problem that is best solved at present through machine learning. However, obtaining large-scale ground-truth data for training deep neural networks remains a challenge, and this has motivated recent work on developing unsupervised schemes for the reflectance separation problem. The unsupervised techniques that have been presented thus far all take the same approach of training a network on image sequences of a fixed scene under changing illumination (Li and Snavely 2018b; Ma et al. 2018). With images from such a sequence, these methods guide network training by exploiting the albedo consistency that exists for each scene point throughout the sequence.

A benefit of using image sequences of fixed scenes is that the images are perfectly aligned, allowing scene point consistency to be easily utilized. However, there exists an untapped wealth of image data captured of objects from different viewpoints. A prominent example of such data is customer product photos uploaded by consumers to show items they bought. Some example customer photos are shown in Figure 1. This source of imagery is valuable not just because of its vast quantity online, but also because it provides object-centric data (different from the scene data compiled in (Li and Snavely 2018b) from webcams) and can promote robustness of factorizations to different object orientations. These images also exhibit a larger variation in illumination conditions and camera settings, which can potentially benefit the trained network. An issue with using such images though is that they are difficult to align accurately, as they vary in viewpoint, lighting and imaging device. Misalignment among the images of an object would lead to violations of scene point consistency on which the existing unsupervised methods are based.

In this paper, we present an unsupervised method for reflectance layer separation using multi-view image sets such as customer product photos. To effectively learn from such data, our system is designed so that its training is relatively insensitive to misalignments. After approximately aligning images with state-of-the-art correspondence estimation techniques (Rocco, Arandjelovic, and Sivic 2018; Ilg et al. 2017), the network transforms the images into a proposed representation based on local color distributions. An important property of this representation is its ability to model detailed local content over an object in a manner that discards fine-scale positional information. With this color distribution based descriptor, unsupervised training becomes possible using consistency constraints between multi-view images of an object.

An additional contribution of this work is a method for further guiding the unsupervised training via a relationship between highlight separation and intrinsic decomposition of

Figure 1: Selected product photos from the Customer Product Photos Dataset. The products exhibit a wide range of textures, shapes, shadings, and highlight patterns. The second last row shows selected multiview images of the same object, where the leftmost one is the segmented reference image. The last row shows the roughly aligned images.

shading and albedo. We observe that shading separation becomes less reliable when highlights are present in its input images, due to color distortions caused by different highlight saturation and possibly different illumination color among the images. Our system takes advantage of this through a novel contrastive loss that is defined between shading separation results computed with and without the inclusion of our highlight extraction sub-network. We show that by maximizing this contrastive loss, the shading separation sub-network provides supervision that improves the performance of the highlight extraction sub-network.

With the presented approach, our system produces state-of-the-art results on highlight separation, and yields intrinsic image decomposition accuracy at a level comparable to leading methods.

Related work

Intrinsic image decomposition Previous to the deep-learning approaches of recent years, intrinsic image decomposition was primarily addressed as an optimization problem constrained by various prior assumptions about natural scenes. These priors have been used to classify image derivatives as either albedo or shading change (Land and McCann 1971; Funt, Drew, and Brockington 1992), to prescribe texture coherence (Shen, Tan, and Lin 2008; Zhao et al. 2012), and to enforce sparsity in the set of albedos (Shen and Yeo 2011; Rother et al. 2011). Decomposition constraints have also been derived using additional input data such as image sequences (Weiss 2001), depth measurements (Lee et al. 2012), and user input (Bousseau, Paris, and Durand 2009).

These earlier methods have been surpassed in performance by deep neural networks which learn statistical priors from training data. Some of these networks are trained with direct supervision, in which the ground-truth albedo and shading components are provided for each training image (Narihira, Maire, and Yu 2015b; Kim et al. 2016; Shi et al. 2017; Baslamisli, Le, and Gevers 2018; Li and Snavely 2018a). To obtain ground truth at a large scale for training deep networks, these methods utilize synthetic renderings, which can lead to poor generalization of the networks to real-world scenes. This issue is avoided in several methods by training on sparse annotations of relative reflectance intensity (Bell, Bala, and Snavely 2014) or relative shading (Kovacs et al. 2017) in real images (Zhou, Krahenbuhl, and Efros 2015; Narihira, Maire, and Yu 2015a; Kovacs et al. 2017; Fan et al. 2018). However, these manual labels provide only weak supervision, and the need for supervision reduces the scalability of the training data.

Most recently, unsupervised methods have been presented in which the training is performed on image sequences taken from fixed-position, time-lapse video with varying illumination (Li and Snavely 2018b; Ma et al. 2018). In these networks, a major source of guidance for unsupervised training is the temporal consistency of reflectance for static regions within a sequence. The networks are configured so that they can be applied to just a single input image at inference time.

Our proposed system also trains on multiple images in an unsupervised manner and can be applied at test time on single images. Different from the previous fixed-view multi-image techniques (Li and Snavely 2018b; Ma et al. 2018), our network uses unconstrained multi-view images and deals specifically with misalignment issues that arise in this setting. Such image sequences from unconstrained random views are much easier to obtain than fixed-view images. Moreover, our method additionally separates highlight reflections and introduces a mechanism by which highlight extraction and intrinsic decomposition can mutually benefit each other in unsupervised training.

We note that multiview images have previously been used for intrinsic image decomposition of outdoor scenes (Laffont, Bousseau, and Drettakis 2013; Duchêne et al. 2015). The decomposition is solved by an inverse rendering approach, where shading is inferred from an approximate multiview stereo reconstruction and an illumination environment estimated given the known sun direction. The multiview images are required to be taken under the same lighting conditions. By contrast, we address a problem where no knowledge about the illumination is given and the lighting can differ from image to image.

Highlight separation Similar to intrinsic image decomposition, separation of highlight reflections is an ill-posed problem that has been made tractable through the use of different priors. Among them are priors on piecewise constancy of surface colors (Klinker, Shafer, and Kanade 1988), smoothness of diffuse (Tan et al. 2003) or specular (Liu et al. 2015) reflection, constancy in the maximum diffuse

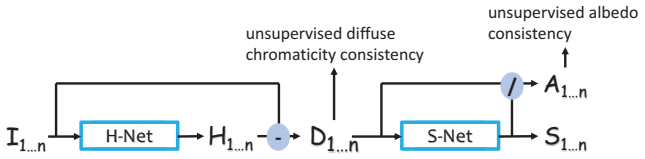

Figure 2: Network structure.

chromaticity (Tan and Ikeuchi 2005), diffuse texture coherence (Tan, Lin, and Quan 2006), low diffuse intensity in a color channel (Kim et al. 2013), sparsity of highlights (Guo, Zhou, and Wang 2018), and a low-rank representation of diffuse reflection (Guo, Zhou, and Wang 2018).

Instead of crafting priors for highlight extraction by hand, they can be learned in a statistical fashion from images using neural networks. This was first investigated together with intrinsic image decomposition through supervised learning on a large collection of rendered images (Shi et al. 2017). An unsupervised approach was later presented for the case of human faces, where a set of images of the same face is aligned using detected facial landmark points, and training guidance is provided by a low-rank constraint on diffuse chromaticity across the aligned images (Yi et al. 2018). In (Yi et al. 2018), face images are easy to align because of mature facial landmark detection techniques; however, their method works poorly on random objects without such landmarks. Thus, we design a much general method to deal with such multi-view images of general objects which are difficult to align accurately. Since misaligned images violate the low-rank property assumed in (Yi et al. 2018), we propose a technique that is robust to such local misalignments, thus enabling unsupervised training over a much broader range of objects. Thus, our method is the first unsupervised method using unconstrained images under random illumination, background, and viewpoints.

Overview

We train an end-to-end deep neural network to separate a single image into highlight, albedo/reflectance, and shading layers using the Customer Product Photos Dataset. Compiled from online shopping websites, the dataset contains numerous product photos provided in customer reviews. The photos for a given product are captured under various viewpoints, illumination conditions, and backgrounds. We introduce this dataset in Section *Customer Product Photos Dataset*.

As illustrated in Figure 2, our network consists of two subnets: H-Net for decomposing an image into diffuse and highlight layers, and S-Net for additionally decomposing the diffuse layer into albedo and shading layers. Training consists of three phases. First, both H-Net and S-Net are pretrained using a small set of synthetic data from ShapeNet (Shi et al. 2017). Each subnet is then finetuned in an unsupervised manner on the Customer Product Photos Dataset using the proposed color distribution loss (Section *Misalignment-robust color distribution loss*), which is robust to misalignments. In the last phase, a novel contrastive loss is used to finetune the whole network end-to-end. The training phases are presented in Section *Our Network*.

Customer Product Photos Dataset

Almost every popular online shopping website includes customer reviews, where customers are often encouraged to upload product photos. For a given product, the customer photos capture it under a various viewpoints, illuminations, and backgrounds. At the same time, the different products cover a large variety of materials and shapes. Collectively, these customer photos capture the complex interaction between different 3D shapes, materials, and illumination, and form a dataset that can be useful for computer vision tasks such as intrinsic image decomposition and multi-view stereo.

Construction of the dataset involved a series of steps consisting of product selection, photo downloading, rough image alignment, and data filtering. Due to limited space, please refer to the supplement[1] for details.

The final Customer Product Photos Dataset consists of 228 products (some shown in Figure 1) with 10–520 photos for each product. In total, the dataset consists of 9,472 photos. For each product, there is one mask provided for the reference image. The original and aligned images will be made available online upon paper publication.

Our Network

Problem formulation

An input image I comprises an additive combination of a highlight layer H and a diffuse layer I_d, where the diffuse layer I_d is a pixelwise product of an albedo/reflectance layer A and a shading layer S, i.e.,

$$I = H + I_d = H + A \cdot S. \tag{1}$$

Our problem is to estimate H, I_d, A, S from the input image I. We note that this image model differs from the conventional intrinsic image model, $I = A \cdot S$, which omits the additive effects of highlights and thus implicitly assumes object surfaces to be matte (Shi et al. 2017).

Low-rank loss for unsupervised training

Most CNN-based methods (Janner et al. 2017; Shi et al. 2017; Narihira, Maire, and Yu 2015a) for intrinsic image separation rely completely on ground truth separation results for supervised training. As it is difficult to obtain reference ground truth for highlight separation or intrinsic image decomposition on real images, we propose to train our network by unsupervised finetuning on real multiview images after an initial supervised pretraining step with synthetic data from the ShapeNet dataset (Shi et al. 2017). This pretraining uses 28,000 out of the 2,443,336 images in the dataset, or about 1.1% of the total, and is intended to provide the network with a good initialization. The finetuning is then intended to adapt the network to the domain of real images, for which ground truth is generally unavailable.

We first assume perfect image alignment in deriving the low-rank loss for unsupervised training. This requirement on alignment will be relaxed in the next subsection.

[1] https://arxiv.org/abs/1911.07262

H-Net For training of highlight separation, our network utilizes input consisting of multiple aligned images $I_1, I_2, I_3, \cdots$ of the same object under different lighting. According to the image formation model, these images each have a diffuse layer, denoted as $I_{d1}, I_{d2}, I_{d3}, \cdots$. These diffuse layers can differ from each other due to changes in shading that arise from different illumination conditions. To discount this shading variation, we compute the chromaticity maps of these diffuse layers. A chromaticity map (Ch_r, Ch_g) is an intensity-normalized image, where

$$Ch_r(p) = \frac{R(p)}{R(p) + G(p) + B(p)},$$

$$Ch_g(p) = \frac{G(p)}{R(p) + G(p) + B(p)},$$

at each pixel p, with $R(p), G(p), B(p)$ denoting the color values at p.

According to the dichromatic reflectance model (Shafer 1985), the chromaticity of diffuse layers is the chromaticity of the surface albedo multiplied with that of the illumination. Assuming a constant illumination color across each image, we discount the effect of illumination chromaticity by matching the median chromaticity of each diffuse image to that of the reference image in each batch. After these normalizations, the set of chromaticity maps should be of low rank if the images are accurately aligned.

The structure of H-Net is adopted from the encoder-decoder network in (Narihira, Maire, and Yu 2015b) with an added batch normalization layer after each convolution layer to aid in network convergence. We also examined adding skip connections between the encoder and decoder as done in (Shi et al. 2017), but we found them not to be helpful in our network.

S-Net Our S-Net for predicting the shading layer S uses the same network structure as H-Net. The albedo layer A is computed from S at each pixel p according to the image formation model, as

$$A(p) = I_d(p)/S(p), \tag{2}$$

once the shading layer is fixed.

For multiple aligned diffuse images $I_{d1}, I_{d2}, I_{d3}, \cdots$ of the same object, their albedo layers $A_1, A_2, A_3, \cdots$ should be the same. Therefore, we can enforce a consistency loss on these different albedo layers for unsupervised training of S-Net.

Low-rank loss Our unsupervised training enforces consistency among diffuse chromaticity layers and albedo layers via a low-rank loss. For the case of albedo layers, the low-rank loss can be defined as the second singular value of the matrix M formed by reshaping each albedo image into a vector and stacking the vectors of multiple images (Yi et al. 2018). Although consistency could alternatively be enforced through minimizing L1 or L2 differences, e.g. minimizing $|A_1 - A_2|_{1,2}$, the lack of scale invariance of the L1 and L2 losses can lead to degenerate results where A_1 and A_2 approach zero. To avoid this problem, the loss function should satisfy the following constraint,

$$\mathcal{L}(A_1, A_2) = \mathcal{L}(\alpha A_1, \alpha A_2),$$

where α is a global scale factor for the whole albedo image.

In order to make the low-rank loss scale-invariant, we use the first singular value to approximate the scale and define a scale-invariant low-rank loss (SILR) as

$$\mathcal{L}_{SILR} = \sigma_2/\sigma_1,$$
$$\frac{\partial \mathcal{L}_{SILR}}{\partial M_{i,j}} = \frac{\sigma_1 * (U_{i,2} \times V_{2,j}) - \sigma_2 * (U_{i,1} \times V_{1,j})}{\sigma_1^2}. \tag{3}$$

where σ_1 and σ_2 are the first two singular value of M computed by SVD decomposition. We apply this scale-invariant low-rank loss (SILR) to train both H-Net and S-Net.

Misalignment-robust color distribution loss

We present a way to relax the requirement of pixel-to-pixel correspondence in the low-rank loss, so that customer photos can be effectively utilized for training. Our observation is that, though precise pixelwise alignment is generally difficult, the state-of-the-art alignment algorithms, e.g. WeakAlign (Rocco, Arandjelovic, and Sivic 2018) and FlowNet (Ilg et al. 2017; Dosovitskiy et al. 2015), are mature enough to establish a reasonable approximate alignment. Thus, though some pixels may be misaligned, their correct correspondences are still within a small neighborhood of their estimated locations. This motivates us to develop a local distribution based representation for the low-rank loss.

Suppose we have a predicted albedo layer A. We partition it into a grid of N cells. Within each cell, we reorder the pixels by increasing intensity. This is done for each color channel individually, and all the cells for all the color channels are reshaped and concatenated to form a new vector representation for the image. The color distribution loss is then computed as the SILR of these image vectors. In our implementation, we divided 320×320 images into 256 grid cells for all training phases.

This vector representation of locally re-ordered pixel values is robust to slight misalignment for the following reasons: (1) Since the dimensions of grid cells are much larger than typical misalignment distances, the corresponding grid cells of different images will largely overlap the same object regions; (2) Products tend to have a sparse set of surface colors, and the pixel reordering will help to align these colors between the corresponding grid cells of different images, which is sufficient for measuring color-based consistency; (3) With this representation, the SILR loss is empirically found to be more sensitive to the presence of highlights or albedo distortions than to slight misalignment, as illustrated in Figure 3 for diffuse chromaticity.

We note that a local color distribution could more directly be modeled by a color histogram. However, color histograms are not differentiable, and this motivated us to develop the pixel reordering representation as a differentiable approximation to color histograms. Local regions that have similar

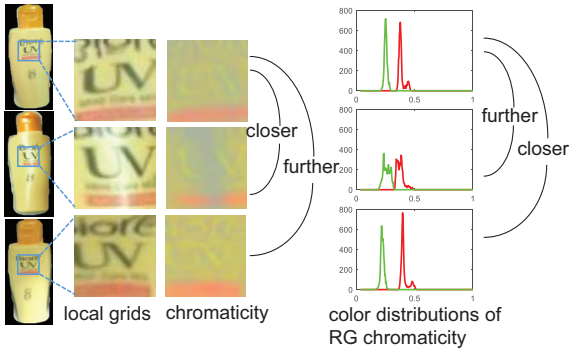

Figure 3: Distances between color distributions are more sensitive to the presence of highlights than to pixel-to-pixel distance between misaligned images. The grid cells in the top two images are spatially closer to each other, but have greater difference in color distribution due to highlights.

color histograms will have similar pixel reordering representations, and vice versa.

Joint finetuning by contrastive loss

After training H-Net and S-Net individually, we adopt a novel contrastive loss to finetune the entire network in an end-to-end manner. Our approach is based on the observation that intrinsic image decomposition can be better performed after highlights have been separated from input images. Related observations have been made in other recent works. For example, Ma et al. (Ma et al. 2018) mention that their method cannot handle specularity well, and this limitation will be addressed in future work. Also, Shi et al. (Shi et al. 2017) discuss that the multiplicative intrinsic image decomposition model, $I_d = A \cdot S$, cannot adequately account for additive highlight components.

Based on this observation, we define a contrastive loss. As indicated in Figure 4, our low-rank loss on the albedo layers of multiple images is $\mathcal{L}_1$ if highlights are removed from the input images following the image formation model $I = A \cdot S + H$. In another branch, we compute the low-rank loss on albedo layers as $\mathcal{L}_0$, where the input images are decomposed by S-Net directly following the image formation model $I = A \cdot S$. The contrastive loss is defined as:

$$\mathcal{L}_{ct} = \mathcal{L}_1 - \mathcal{L}_0. \tag{4}$$

Intuitively, the contrastive loss is designed to maximize the distance between $\mathcal{L}_1$ and $\mathcal{L}_0$ (where $\mathcal{L}_{ct}$ is negative), so as to force H-Net to improve its highlight separation and thus decrease $\mathcal{L}_1$ relative to $\mathcal{L}_0$. Both subnets can be finetuned by this loss. In our experiments, we found that using $\mathcal{L}_{ct}$ alone will lead to increases of both $\mathcal{L}_1$ and $\mathcal{L}_0$, as this increases their difference as well. To avoid this degenerate case, we add $\omega\mathcal{L}_1$ as a regularization, such that the joint finetuning loss becomes $\mathcal{L} = \mathcal{L}_{ct} + \omega\mathcal{L}_1$, where ω is set to 1.0 in our implementation. This ensures that both $\mathcal{L}_1$ and the contrastive loss are minimized together.

After these three training phases, our network shown in Figure 2 is able to separate the highlight, diffuse, albedo, and shading layers of a test image. Further implementation details are given in the supplement[1].

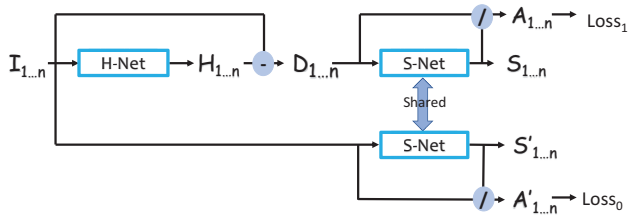

Figure 4: Network structure for joint finetuning by contrastive loss.

Experiments

Since previous works generally address highlight separation or intrinsic image estimation but not both, we evaluate our method on these two tasks separately on various datasets. Due to limited space, many additional results and analyses, including evaluations on the MIT intrinsic image dataset (Grosse et al. 2009) and Intrinsic Images in the Wild (IIW) (Bell, Bala, and Snavely 2014), highlight separation on grayscale images (which cannot be handled by most previous techniques), and the inadequacy of structure-from-motion for aligning our customer photos, are provided in the supplement[1].

Highlight separation

Synthetic dataset In Table 1 (top-left), we compare our method to several leading techniques on highlight separation using synthetic data from the ShapeNet Intrinsic Dataset (Shi et al. 2017). From this dataset, we randomly select 500 images covering a wide range of objects and materials to form the test set. The results are reported in terms of MSE and DSSIM, which measure pixelwise difference and structural dissimilarities, respectively.

Examples for visual comparison are shown in Figure 5. Earlier methods (Tan, Nishino, and Ikeuchi 2004; Yang, Wang, and Ahuja 2010; Shen and Zheng 2013) often assume the illumination to be white and can estimate only a grayscale highlight layer, even when the lighting is not white. Moreover, they cannot deal well with saturated regions, which generally have non-white highlight components that result from subtracting (non-white) diffuse components from saturated image values. A recent method (Guo, Zhou, and Wang 2018) handles saturated highlight regions better with a low-rank and sparse decomposition. However, it still cannot recover correct diffuse color at saturated regions where its assumed dichromatic model is violated, leading to artifacts in diffuse layers. The CNN-based method of (Shi et al. 2017) can learn from various training data composed of different surface materials, but it still does not handle saturation well. By comparison, our method succeeds in predicting highlight colors and generates reasonable diffuse layers even for saturated regions.

Real dataset Since no standard real-image dataset exists for evaluating highlight separation, we captured a dataset

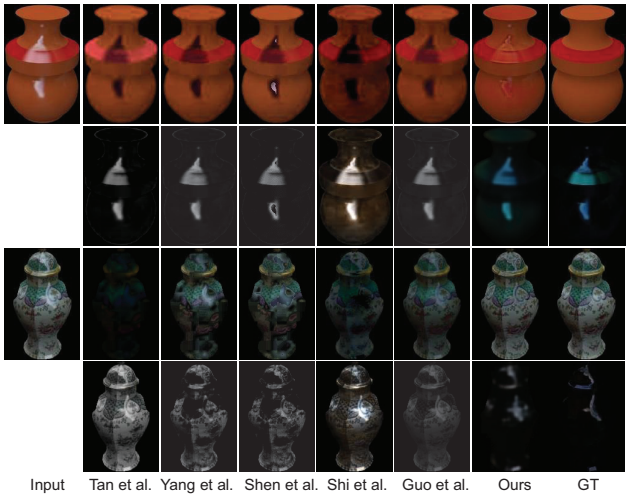

Figure 5: Visual comparisons of highlight separation on the ShapeNet Intrinsic Dataset. For each example, the top row shows the input image and separated diffuse layers, and the bottom row exhibits the separated highlight layers. GT denotes ground truth.

Method	Synthetic		Real	
	MSE	DSSIM	MSE	DSSIM
Tan et al.	0.0155	0.0616	0.0173	0.0368
Yang et al.	0.0053	0.0336	0.0043	0.0162
Shen et al.	0.0059	0.0338	0.0047	0.0163
Shi et al.	0.0063	0.0526	0.0063	0.0237
Guo et al.	0.0028*	0.0208*	0.0045	*0.0145*
Ours	*0.0016*	**0.0159**	**0.0036**	**0.0139**
No Finetuning	**0.0015**	0.0176	0.0045	0.0188
Pixel-to-pixel	0.0020	*0.0166*	*0.0041*	0.0149

Table 1: Highlight separation on the synthetic ShapeNet Intrinsic Dataset and on a real-image dataset. Errors are for diffuse layers. **Top:** Comparison to state of the art. Lowest errors shown in bold and second lowest in italics. Guo(Guo, Zhou, and Wang 2018) is tested on 50 of the 500 synthetic data in total, with the results marked by *, since we needed the authors to process our images. **Bottom:** Ablations.

consisting of 20 ordinary objects with ground truth obtained by cross polarization in a laboratory environment. Table 1 (top-right) shows the MSE and DSSIM of different methods on this dataset. Qualitative comparisons are shown in Figure A.6 and A.7 of the supplement[1]. Our method is found to recover highlight and diffuse layers closest to the ground truth, with highlights of correct color even in saturated regions. While our technique successfully estimates the surface colors in the diffuse layers, the other methods tend to leave black artifacts at saturated regions. Additional qualitative results on real images under natural lighting can be found in the supplement[1] as well.

Ablations We conducted an ablation study to examine the main novel elements of our system, with the results shown in Table 1 (bottom). When the unsupervised finetuning is removed from the system, the difference in performance becomes more significant on real images than on synthetic images, since the finetuning provides training in the domain of real images. On real images without finetuning, the performance is at a level similar to the previous state of the art, while our full system yields an approximate 20-25% improvement over this.

To examine the importance of our color distribution loss in dealing with misalignment, we compare to the results of our network when using a pixel-to-pixel low-rank loss instead. Some moderate quantitative gain is observed, about 4-20% for synthetic images and 7-10% for real images. We point readers to the qualitative comparisons shown in Fig. A.1 of the supplement[1], where the diffuse layers computed without the color distribution loss contain severe artifacts around highlight regions. Later, it will be shown that the color distribution loss has greater quantitative impact on intrinsic image decomposition.

When the contrastive loss is removed from the system, the solution often degenerates to a diffuse layer of all zeros, as this allows H-Net to reach a minimum most quickly. Similar to a generative adversarial network (GAN), the contrastive loss creates a competition between losses that can steer the learning toward better minima and/or away from degenerate cases. By including the contrastive loss, the learning rate of S-Net becomes twice that of H-Net, causing the training to focus more on S-Net and thus avoiding degenerate solutions.

	MSE(A)	DSSIM(A)	MSE(S)	DSSIM(S)
SIRFS	0.0081	0.0636	0.0066	0.0785
DI	0.0086	0.0590	0.0047	0.0765
Shi et al.	0.0068	0.0565	**0.0023**	*0.0691*
Li et al.	*0.0066*	0.0541	0.0063	0.0812
Ours	**0.0054**	**0.0436**	*0.0045*	**0.0686**
No Finetuning	0.0108	0.0664	0.0096	0.0810
Pixel-to-pixel	0.0067	*0.0460*	0.0087	0.0774

Table 2: Intrinsic image decomposition on synthetic data from the ShapeNet Intrinsic Dataset. The lowest errors are in bold, and the second lowest are in italics.

Intrinsic image decomposition

ShapeNet Intrinsic Dataset For intrinsic image decomposition, we compare our network to SIRFS (Barron and Malik 2015), DI (Narihira, Maire, and Yu 2015b), Shi et al. (Shi et al. 2017), and Li et al. (Li and Snavely 2018b) on the ShapeNet Intrinsic Dataset. Similar to the evaluation of highlight separation, we use MSE and DSSIM to measure results. These results are summarized in Table 2 (top) and show the relatively strong performance of our method. Qualitative comparisons are shown in Figure A.13 and A.14 of the supplement[1].

SIRFS (Barron and Malik 2015), which is based on scene priors, fails on non-Lambertian objects. The learning-based method DI (Narihira, Maire, and Yu 2015b) trained on synthetic diffuse scenes exhibits similar problems. The method by Shi et al. (Shi et al. 2017) performs better than previous methods on non-Lambertian objects. One reason is that, like our method, it explicitly models highlights, in contrast to other methods (Narihira, Maire, and Yu 2015b; Barron and Malik 2015; Li and Snavely 2018b) which consequently

	MSE(S)	DSSIM(S)
SIRFS(Barron and Malik 2015)	0.0097	0.0457
DI(Narihira, Maire, and Yu 2015b)	0.0061	0.0385
Shi(Shi et al. 2017)	0.0043	0.0331
Li(Li and Snavely 2018b)	0.0073	0.0401
CG(Li and Snavely 2018a)	0.0061	0.0413
Ours	**0.0041**	**0.0316**

Table 3: Evaluation of shading accuracy on the DiLiGenT dataset. The lowest errors are in bold.

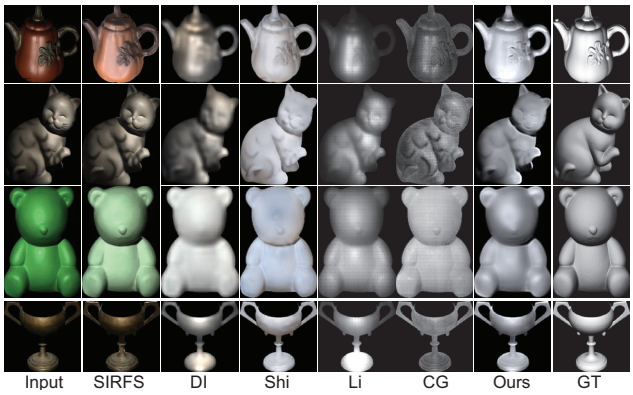

Figure 6: Shading layer comparisons on DiLiGenT dataset. Please see Table 3 for the notations of previous methods.

have artifacts in the albedo layer on highlight regions. Another reason is because it is trained on the ShapeNet Intrinsic training split with 80% of the whole dataset. In comparison, our method is pretrained on a very small amount (1.1%) of the ShapeNet dataset to obtain a good network initialization, and is finetuned on a large amount of real data. Despite this, it still performs well on synthetic ShapeNet images. Since our S-Net solves for shading and then computes albedo using the image formation model $I_d = A \cdot S$, it generates high resolution albedo maps with texture details, whereas many networks that directly solve for albedo will obtain blurred results due to feature map downsampling in the network.

DiLiGenT dataset We also conduct experiments on real images. Since there do not exist intrinsic image datasets with ground truth for general real objects[2], we evaluate on ground-truth shading layers generated from the DiLiGenT photometric stereo dataset (Shi et al. 2019). As DiLiGenT provides ground-truth surface normals and lighting, but no reflectance information, only the shading layers can be reconstructed. The dataset contains images of 10 non-Lambertian objects under 96 different lighting conditions.

Comparisons of our network are made to several leading techniques. Qualitative and quantitative results are shown in Figure 6 and Table 3. It is found that our network yields the highest accuracy in this challenging case of real non-Lambertian objects.

[2]The IIW dataset (Bell, Bala, and Snavely 2014) and SAW dataset (Kovacs et al. 2017) are of real *scenes*, while the objects in the MIT dataset (Grosse et al. 2009) are restricted to highly Lambertian reflectance.

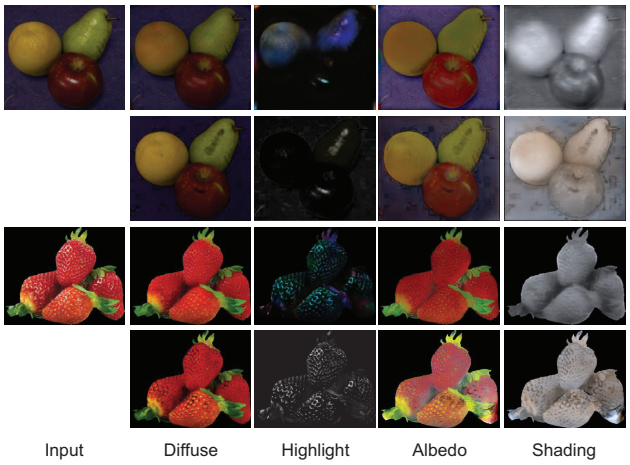

Figure 7: Qualitative comparisons on real images. We compare our end-to-end separation (odd rows) to the combination of Yang (Yang, Wang, and Ahuja 2010) for highlight separation and Shi (Shi et al. 2017) for intrinsic image decomposition (even rows).

Other datasets There exist other datasets that can be used for intrinsic image evaluation, including the MIT intrinsic image dataset (Grosse et al. 2009) and Intrinsic Images in the Wild (IIW) (Bell, Bala, and Snavely 2014). Due to limited space, comparisons on these datasets, as well as qualitative comparisons on more natural images collected from the Internet, are presented in the supplement[1]. In addition, some qualitative results of full end-to-end separations on real images are shown in Figure 7, with comparisons to a combination of two previous methods that exhibit state-of-the-art performance in quantitative evaluations.

Ablations Ablation experiments were also conducted for intrinsic image decomposition on ShapeNet, with the results given in Table 2 (bottom). Even though ShapeNet consists of synthetic images, significant gains were obtained by including the unsupervised finetuning (15-50%) and by using the color distribution loss instead of a pixel-to-pixel low rank loss (5-48%). The difference is particularly large for shading, as also evidenced in the qualitative comparisons shown in Figure A.1 of the supplement[1] where the shading layers are more indicative of surface shape. As with highlight separation, removal of the contrastive loss leads to degenerate solutions where the diffuse layer is all zero.

Conclusion

We proposed an end-to-end network to solve highlight separation and intrinsic image decomposition together. Our network is able to leverage multi-view object-centric image sets, such as our Customer Product Photos Dataset, for unsupervised training via a proposed color distribution loss that is robust to misaligned data. This loss can readily be adapted for other tasks that are sensitive to misalignment.

Acknowledgments. This work is supported by Canada NSERC Discovery Grant 611664.

References

Barron, J. T., and Malik, J. 2015. Shape, illumination, and reflectance from shading. *IEEE TPAMI* 37(8):1670–1687.

Baslamisli, A. S.; Le, H.-A.; and Gevers, T. 2018. Cnn based learning using reflection and retinex models for intrinsic image decomposition. In *ICCV*.

Bell, S.; Bala, K.; and Snavely, N. 2014. Intrinsic images in the wild. *ACM Transactions on Graphics* 33(4).

Bousseau, A.; Paris, S.; and Durand, F. 2009. User-Assisted Intrinsic Images. *ACM Trans. Graph* 28.

Dosovitskiy, A.; Fischer, P.; Ilg, E.; Hausser, P.; Hazirbas, C.; Golkov, V.; Van Der Smagt, P.; Cremers, D.; and Brox, T. 2015. Flownet: Learning optical flow with convolutional networks. In *ICCV*, 2758–2766.

Duchêne, S.; Riant, C.; Chaurasia, G.; Moreno, J. L.; Laffont, P.-Y.; Popov, S.; Bousseau, A.; and Drettakis, G. 2015. Multiview intrinsic images of outdoors scenes with an application to relighting. *ACM Trans. Graph.* 34(5):164:1–164:16.

Fan, Q.; Yang, J.; Hua, G.; Chen, B.; and Wipf, D. 2018. Revisiting deep intrinsic image decompositions. In *CVPR*.

Funt, B. V.; Drew, M. S.; and Brockington, M. 1992. Recovering shading from color images. In *ECCV*, 124–132.

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

Guo, J.; Zhou, Z.; and Wang, L. 2018. Single image highlight removal with a sparse and low-rank reflection model. In *ECCV*.

Ilg, E.; Mayer, N.; Saikia, T.; Keuper, M.; Dosovitskiy, A.; and Brox, T. 2017. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *CVPR*.

Janner, M.; Wu, J.; Kulkarni, T. D.; Yildirim, I.; and Tenenbaum, J. 2017. Self-supervised intrinsic image decomposition. In *NIPS*, 5936–5946.

Jeon, J.; Cho, S.; Tong, X.; and Lee, S. 2014. Intrinsic image decomposition using structure-texture separation and surface normals. In *ECCV*.

Kim, H.; Jin, H.; Hadap, S.; and Kweon, I. 2013. Specular reflection separation using dark channel prior. In *CVPR*, 1460–1467.

Kim, S.; Park, K.; Sohn, K.; and Lin, S. 2016. Unified depth prediction and intrinsic image decomposition from a single image via joint convolutional neural fields. In *ECCV*.

Klinker, G.; Shafer, S.; and Kanade, T. 1988. The measurement of highlights in color images. *IJCV* 2(1):7–32.

Kovacs, B.; Bell, S.; Snavely, N.; and Bala, K. 2017. Shading annotations in the wild. In *CVPR*.

Laffont, P.-Y.; Bousseau, A.; and Drettakis, G. 2013. Rich intrinsic image decomposition of outdoor scenes from multiple views. *IEEE TVCG* 19(2):210–224.

Land, E., and McCann, J. 1971. Lightness and retinex theory. *Journal of the Optical Society of America* 3:1684 – 1692.

Lee, K. J.; Zhao, Q.; Tong, X.; Gong, M.; Izadi, S.; Lee, S. U.; Tan, P.; and Lin, S. 2012. Estimation of intrinsic image sequences from image+depth video. In *ECCV*, 327–340.

Li, Z., and Snavely, N. 2018a. Cgintrinsics: Better intrinsic image decomposition through physically-based rendering. In *ECCV*.

Li, Z., and Snavely, N. 2018b. Learning intrinsic image decomposition from watching the world. In *CVPR*.

Li, C.; Zhou, K.; and Lin, S. 2015. Simulating makeup through physics-based manipulation of intrinsic image layers. In *CVPR*.

Liu, Y.; Yuan, Z.; Zheng, N.; and Wu, Y. 2015. Saturation-preserving specular reflection separation. In *CVPR*.

Lombardi, S., and Nishino, K. 2016. Reflectance and illumination recovery in the wild. *IEEE TPAMI* 38(1):129–141.

Ma, W.-C.; Chu, H.; Zhou, B.; Urtasun, R.; and Torralba, A. 2018. Single image intrinsic decomposition without a single intrinsic image. In *ECCV*, 211–229.

Narihira, T.; Maire, M.; and Yu, S. X. 2015a. Learning lightness from human judgement on relative reflectance. In *CVPR*.

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

Or-El, R.; Rosman, G.; Wetzler, A.; Kimmel, R.; and Bruckstein, A. M. 2015. Rgbd-fusion: Real-time high precision depth recovery. In *CVPR*.

Rocco, I.; Arandjelovic, R.; and Sivic, J. 2018. End-to-end weakly-supervised semantic alignment. In *CVPR*.

Rother, C.; Kiefel, M.; Zhang, L.; Scholkopf, B.; and Gehler, P. V. 2011. Recovering intrinsic images with a global sparsity prior on reflectance. In *NIPS*, 765–773.

Shafer, S. A. 1985. Using color to separate reflection components. *Color Research & Application* 10(4):210–218.

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

Shen, H.-L., and Zheng, Z.-H. 2013. Real-time highlight removal using intensity ratio. *Applied optics* 52(19):4483–4493.

Shen, L.; Tan, P.; and Lin, S. 2008. Intrinsic image decomposition with non-local texture cues. In *CVPR*.

Shi, J.; Dong, Y.; Su, H.; and Yu, S. X. 2017. Learning non-lambertian object intrinsics across shapenet categories. In *CVPR*, 5844–5853.

Shi, B.; Mo, Z.; Wu, Z.; Duan, D.; Yeung, S.-K.; and Tan, P. 2019. A benchmark dataset and evaluation for non-lambertian and uncalibrated photometric stereo. *IEEE TPAMI* 41(2):271–284.

Tan, R. T., and Ikeuchi, K. 2005. Separating reflection components of textured surfaces using a single image. *IEEE TPAMI* 27(12):178–193.

Tan, P.; Lin, S.; Quan, L.; and Shum, H.-Y. 2003. Highlight removal by illumination-constrained inpainting. In *ICCV*.

Tan, P.; Lin, S.; and Quan, L. 2006. Separation of highlight reflections on textured surfaces. In *CVPR*, volume 2, 1855–1860.

Tan, R. T.; Nishino, K.; and Ikeuchi, K. 2004. Separating reflection components based on chromaticity and noise analysis. *IEEE TPAMI* 26(10):1373–1379.

Weiss, Y. 2001. Deriving intrinsic images from image sequences. In *ICCV*, 68–75.

Yang, Q.; Wang, S.; and Ahuja, N. 2010. Real-time specular highlight removal using bilateral filtering. In *ECCV*, 87–100.

Yi, R.; Zhu, C.; Tan, P.; and Lin, S. 2018. Faces as lighting probes via unsupervised deep highlight extraction. In *ECCV*.

Yu, L.-F.; Yeung, S.-K.; Tai, Y.-W.; and Lin, S. 2013. Shading-based shape refinement of rgb-d images. In *CVPR*.

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 TPAMI* 34(7):1437–1444.

Zhou, T.; Krahenbuhl, P.; and Efros, A. A. 2015. Learning data-driven reflectance priors for intrinsic image decomposition. In *ICCV*.