Pacific Graphics 2020
E. Eisemann, A. Jacobson, and F.-L Zhang
(Guest Editors)

Volume 39 (2020), Number 7

Deep Separation of Direct and Global Components from a Single Photograph under Structured Lighting

Z. Duan J. Bieron P. Peers

College of William & Mary

Abstract

We present a deep learning based solution for separating the direct and global light transport components from a single photograph captured under high frequency structured lighting with a co-axial projector-camera setup. We employ an architecture with one encoder and two decoders that shares information between the encoder and the decoders, as well as between both decoders to ensure a consistent decomposition between both light transport components. Furthermore, our deep learning separation approach does not require binary structured illumination, allowing us to utilize the full resolution capabilities of the projector. Consequently, our deep separation network is able to achieve high fidelity decompositions for lighting frequency sensitive features such as subsurface scattering and specular reflections. We evaluate and demonstrate our direct and global separation method on a wide variety of synthetic and captured scenes.

Keywords: Direct-global Separation, Single Photograph, CNN

1. Introduction

Many inverse methods in computer graphics assume that direct lighting dominates the observations. The presence of significant global light transport breaks this key assumption, often adversely affecting the accuracy of such inverse algorithms. In seminal work Nayar et al. [NKGR06] introduced a method for separating direct and global reflectance from a few measurements of the scene lit by high frequency shifted stripe, checkerboard, or sine wave lighting patterns. Nayar et al.'s key insight is that, in contrast to direct reflectance, the resulting global reflectance is approximately invariant under shifted high frequency lighting. However, Nayar et al.'s method requires multiple photographs of the scene, and is therefore only suited for static scenes. Single-photograph decomposition has been introduced to overcome the inherent limitations of multi-photograph methods. However, separating direct and global lighting from a single photograph is an underconstrained problem. Existing single-photograph solutions either rely on specialized hardware [ORK12, OMK16], or sacrifice spatial resolution or sharpness [NKGR06, SaFZ*18].

The majority of single-photograph and multi-photograph methods rely on high frequency illumination to introduce additional cues to help in decomposing the direct and global components. The higher frequency the lighting patterns, the higher the granularity at which direct and global lighting can be disentangled. This becomes particularly important for light paths that vary quickly in either the direct or global component such as specular reflections, caustics, and single scattering (vs. subsurface scattering) light transport. The ready availability of low cost projectors makes them a convenient tool to induce such high frequency lighting. However, most methods assume binary lighting patterns, and typical low-cost consumer-grade projectors cannot produce *sharp* high contrast binary patterns at their highest resolution, necessitating lower resolutions to mitigate the effects of blurring. While reducing the resolution improves robustness, it also adversely affects the ability to accurately separate lighting frequency sensitive indirect transport effects such as subsurface scattering and specular reflections.

In this paper we propose a novel method for decomposing the direct and global components of a scene that aims to both *maximize the frequency content of the lighting*, thereby improving accuracy for decomposing lighting frequency sensitive effects, and at the same time only requiring *a single photograph* of the scene taken with a co-axial camera-projector system. Our method relies on deep learning, and we employ two identical encoder-decoder networks, with shared encoder weights, that jointly regress the direct and global components at the *full projector resolution*. To promote consistency between both decoders we not only employ skip connections between the encoder and decoder, but also between the direct and global decoders. We illuminate the scene with structured lighting using the *full* available resolution of consumer-grade projectors to resolve the ambiguities inherent in decomposing an image in its direct and global components. Projecting high frequency lighting patterns at full projector resolution is practically challenging, as consumer-grade projectors introduce many types of spatial artifacts in the projected illumination due to compromises in the optical system or due to 'clever' firmware-based image enhancements. As a result, we cannot assume that binary patterns remain binary after projection, and the artifacts introduced need to be taken

into account during training in order to make the deep network robust to typical acquisition conditions.

Key to our method is that it is designed from the start with *imperfect non-binary* structured lighting in mind. To tailor the network to the specific imperfections of an acquisition setup, we refrain from adding additional ad-hoc loss function terms or requiring the arduous acquisition of an extensive training set for each new capture setup, but instead we leverage a light weight method for adapting a small publicly available general training set to the peculiarities of the capture setup that only requires two photographs of a white planar surface captured with the target setup. In addition, we use an effective data-enhancement method to further augment the small set of 20 publicly available training decompositions to ensure that our method generalizes to a wide variety of scenes and materials.

Typically, the resolution of consumer cameras exceeds that of projectors. Ideally we would like to separate the lighting components at camera resolution instead of at projector resolution. However, at camera resolution, the variations in lighting are blurred and the network has difficulty distinguishing scene edges from lighting edges. To address this issue, we augment our deep direct-global separation network training with a novel total variation loss-term variant that takes into account the structured lighting. This allows us to decompose photographs at native camera resolution while maintaining sharp scene features.

In summary, we propose a deep learning based direct-global decomposition method that:

- only requires a single input photograph of a scene under a fixed high frequency lighting pattern;
- uses the full projector resolution, enabling more accurate decompositions compared to other existing single and, in some cases, multiple image decomposition methods for high frequency light paths;
- only requires two calibration images per hardware setup suited for off-the-shelf projectors;
- avoids additional ad-hoc loss terms or dedicated training data acquisition steps, but instead relies on a novel low-weight training augmentation scheme based on a small set of publicly available direct-global decompositions to adapt the neural network to the idiosyncrasies of the capture setup; and
- that can be extended to handle sharp full camera resolution decompositions through a novel total variation loss function.

We demonstrate our method on a variety of scenes, and perform a careful analysis of the capabilities and limitations on synthetic and real-world test cases.

2. Related Work

Direct-Global Component Separation Nayar et al. [NKGR06] formulate image formation of a scene lit by a digital projector as the sum of a direct reflectance component $\mathbf{I_d}$ that is directly modulated by the projector lighting $\mathbf{L}$, and a global reflectance component $\mathbf{I_g}$ that encompasses the resulting reflectance from all indirect lighting incident at the corresponding surface point:

$$\mathbf{I} = \mathbf{I_d}\mathbf{L} + \mathbf{I_g} w(\mathbf{L}), \tag{1}$$

where $w(\mathbf{L})$ represents the ratio of indirect reflectance produced under the lighting $\mathbf{L}$ with respect to the indirect reflectance under uniform white lighting. Nayar et al.'s key insight is that under high frequency illumination (i.e., that varies rapidly over neighboring pixels), $w(\mathbf{L})$ is approximately constant. By further constraining the lighting to be binary and lighting half the projector pixels, the resulting $w(\mathbf{L}) = 0.5$. Under binary lighting, the direct component will either be visible (lit projector pixel) or not (unlit projector pixel). In theory, by capturing the scene under two complementary high frequency projector lighting patterns, the direct $\mathbf{I_d}$ and global component $\mathbf{I_g}$ can be computed. Due to practical limitations of projector systems (e.g., defocus, edge blurring, etc.) and the camera (e.g., camera noise), multiple observations under shifted high frequency illumination are needed to robustly solve Equation 1.

Subsequent work extended direct-global component separation to multiple light sources [GKGN11], to compensate for motion [ANN13], to high-speed capture [KYN12], to real-time visualization of indirect-only images [OANK15], to overcome limitations in depth of field [GTNZ12, AN14], and to further refine the global component in near and far-range transport [RRC12]. Other work has explored more exotic hardware such as time-of-flight cameras [WVO*14, OHX*14], and by adding an additional controllable mask in front of the camera [ORK12, OMK16].

All of the above methods either rely on multiple images and/or complex hardware and/or can only extract a single component at the time. Our method uses a more common co-axial camera-projector setup and only requires a single photograph to extract both the direct and global components. Furthermore, all of the above methods rely on explicit calibration and/or an overcomplete set of lighting patterns to compensate for acquisition hardware deficiencies. In contrast, we take a data-driven approach based on deep learning and convolutional neural networks that are trained to take hardware specific artifacts automatically into account.

Single Image Direct-Global Component Separation In their seminal work, Nayar et al. [NKGR06] also propose a single image solution by searching for the minimum and maximum pixel values in a small window, and then interpolating the resulting direct and global components to the desired output resolution. Subpa et al. [SaFZ*18] perform direct-global separation from a single image assuming sparsity in a Fourier or PCA basis, and assuming some spatial smoothness. Both methods rely on binary patterns, typically displayed at a $4\times$ or $8\times$ lower resolution than the full projector resolution to avoid display artifacts. This has two direct consequences. First, using a lower effective resolution acts as a low-pass cut-off on the light transport that can be separated, effectively incorrectly assigning some indirect lighting to the direct component. Depending on the lighting effect (e.g., specular reflections and subsurface scattering), this error can be significant. Second, the direct component for camera pixels that are not directly lit need to be inferred from the indirect lighting and/or by interpolation. Lower resolution *binary* lighting requires more indirect inference or longer range interpolation, and thus can introduce additional errors in the direct component. In contrast, our learning based separation method can leverage the full projector resolution ensuring a more accurate separation of high frequency light transport effects. Furthermore, be-

cause we rely on non-binary patterns, each captured pixel provides some information on both the direct and indirect component.

Recently, Nie et al. [NGS*18] introduced a cycleGAN [ZPIE17] inspired image-translation solution with a fixed inverse mapping to decompose a single photograph, trained on a large (100) training set of direct-global decompositions under uniform white lighting. Nie et al. show that their method generalizes well to general uncontrolled lighting. While more flexible and independent of the lighting frequency, a decomposition under uncontrolled lighting is a more ill-conditioned separation problem as there is no direct observation of the differences of both components. The use of non-binary lighting patterns in our solution strikes a balance between observing spatial image detail in both components and observing the differences between the direct and global components.

3. Deep Separation of Direct and Global Components

We endeavor to separate the direct and global components from a single photograph. This is an ill-conditioned problem as there are more unknowns ($\mathbf{I_d}$ and $\mathbf{I_g}$) then knowns ($\mathbf{I}$ in Equation 1). To keep the problem tractable, we will assume that the *controlled* incident lighting $\mathbf{L}$ is *exactly* known. For convenience we also assume that all components (direct $\mathbf{I_d}$ and global $\mathbf{I_g}$), the photograph $\mathbf{I}$, and the incident lighting $\mathbf{L}$ are all expressed in the same parameterization. Since projectors typically have a lower resolution, we will assume that all images are parameterized according to the projector's resolution and 'view' of the scene; in section 6 we will relax the constraint that the camera and projector resolution are identical. We will also first explain our algorithm for general full projector resolution lighting patterns; section 4 details the specific high frequency pattern we use in practice.

Network Structure We use two regular fully convolutional encoder-decoder networks to separate the direct and global components, and share the encoder weights between both networks. The structure of the direct and global components are closely related to the structures in the input photograph, as well as to each other. We therefore include skip connections between the different layers of the encoder and decoder, as well as cross connections between the both decoders to promote sharing the embedded relations.

The goal of the network is to encode the relations between observed reflectance under structured lighting in a local window to a clean direct and global component without spatial variations due to the structured lighting. Note, even though global light transport is a non-local light transport phenomena, as observed by Nayar et al. [NKGR06], under high frequency lighting the effect of global lighting can be encoded locally. We therefore, train our network on 128×128 patches; working with small patches limits the size of the network and simplifies training. However, using structured lighting also implies that the incident lighting will vary over the image, and thus between patches (e.g., due to projector fall-off, chromatic aberrations, etc.). We therefore also provide the corresponding 128×128 patch from the incident lighting. Practically, we concatenate the incident lighting $\mathbf{L}$ to the captured image $\mathbf{I}$ to form a 6-channel 128×128 input image. Figure 1 summarizes our network structure.

During inference, we support larger image resolution than 128×128 by exploiting the fully convolutional nature of our network architecture, and expand the size of the bottleneck proportionally to the size of the input photographs.

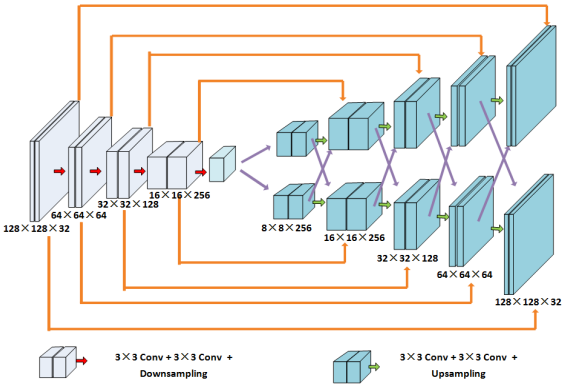

Figure 1: *Deep direct-global separation network consisting of a shared encoder and dual decoder branches connected to the encoder via skip connections, as well as cross connections between both decoders.*

Loss Function To train our separation network, we employ a regular loss function that directly measures the accuracy of the direct component, and the global component:

$$E_{sep} = \mathcal{M}(\widehat{\mathbf{I_d}}, \mathbf{I_d}) + \mathcal{M}(\widehat{\mathbf{I_g}}, \mathbf{I_g}), \tag{2}$$

where $\widehat{\mathbf{I_d}}$ and $\widehat{\mathbf{I_g}}$ are the reference direct and global component respectively, and $\mathcal{M}(\cdot,\cdot)$ is a distance metric. We purposefully kept our loss function simple, refraining from including ad hoc terms to bias the solution, and instead we aim to include all relevant learnable features in our training data.

We have experimented with a number of different distance metrics (subsection 5.5), and found that computing the L_2 distance in LAB color space produced the sharpest results with best color fidelity. Note that we store the components in RGB and only convert to LAB for computing the distance. Storing the components in LAB would make Equation 1 non-linear and thus more difficult to regress. We also experimented with adding a sum-constraint $\mathcal{M}(\widehat{\mathbf{I_d} + \mathbf{I_g}}, \mathbf{I_d} + \mathbf{I_g})$ as in Nie et al. but found that this neither improved nor degraded the quality of the separation.

Training Data Due to the simplicity of our loss function, the training dataset needs to be sufficiently varied and rich to encode all relevant features necessary to successfully decompose the components and generalize beyond the training data.

A critical design decision that needs to be addressed before designing the training data is whether or not to make the network invariant to the optical characteristics of the setup. Training an setup-agnostic network requires a more rich training set that includes all possible variations introduced by different setups. While such an agnostic network is more flexible, it can lead to reduced accuracy especially for low signal-to-noise capture setups that employ commodity projectors. Training a setup-specific network on the other hand, requires training data that embeds the optical artifacts and characteristics of the setup. The most straightforward way to obtain

such a training set is to capture one with the target setup. However, this approach is labor intensive and cumbersome, and it introduces uncertainty on whether the captured scenes are sufficiently rich to ensure successful training. To get the best of both, we will train a setup-specific network, but instead of capturing training data for each setup, we will synthesize setup-specific exemplars.

A common strategy is to generate such synthetic images using a global illumination rendering algorithm. However, we found that they are not suited for training our network. A key issue is that it is very easy to compute a clean, noise free, direct image, while it is much more difficult to obtain an equally clean and noise free indirect image. Even for very high sample rates, there is still some minute amount of noise, and consequently, a trained network tends to assign noise and image details to the global image. Instead, assuming knowledge of the lighting pattern $\mathbf{L}$ as emitted by the setup (section 4), we will leverage Equation 1 and simulate the capture from decompositions of a set of captured reference scenes. Given a reference direct-global decomposition patch of 128×128 resolution, we select a random 128×128 patch from the target lighting pattern $\mathbf{L}$, and simulate the resulting acquisition according to Equation 1. We compute $w(\mathbf{L})$ as the average over a 20×20 window over $\mathbf{L}$. While Equation 1 is only an approximation, we found that the resulting network generalizes well to real-world acquired photographs.

Ideally, the set of reference decomposition should be sufficiently rich and representative. Unfortunately, there currently does not exist a sufficiently large publicly available dataset of direct-global reference decompositions, and only a few (≈ 20) high quality decompositions are publicly available [NKGR06, AN14]. To facilitate reproduction, we base our training set on this publicly available small set of 20 direct-global decompositions (Figure 2), and apply an extensive set of augmentations to enrich the training data. We apply the following augmentations on the reference direct and global components simultaneously:

1. Select a random 128×128 patch from the training images. Note that the images in the training set are significantly larger than 128×128, and thus a large variety of patches can be sampled. We discard patches for which the mean of both the direct and global component falls below 0.003 as these represent patches with little information;
2. Randomly rotate each patch by $0, 90, 180$, or 270 degrees;
3. Upsample each patch by a random scale factor $\in [1,2]$, and crop a 128×128 patch;
4. Randomly flip the patch vertically and/or horizontally;
5. Consider all 6 permutations of color channels (i.e., $rgb, rbg, grb, gbr, brg, bgr$);
6. Scale each color channel by a random scale factor $\in [0,1]$ with a probability of 95%; we keep the original color channel scaling for the other 5%.

We apply the above augmentation online during training, and thus each epoch a different set of $170,610$ training samples (and $48,000$ test exemplars) is used. We found the color channel swapping and color scaling augmentation to be the most critical for promoting generalization to scenes that differ significantly from the training set.

Implementation We implemented our network in PyTorch. We train the network for 80 epochs using Adam [KB14] with the following hyper-parameters: $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a learning rate of 0.0001.

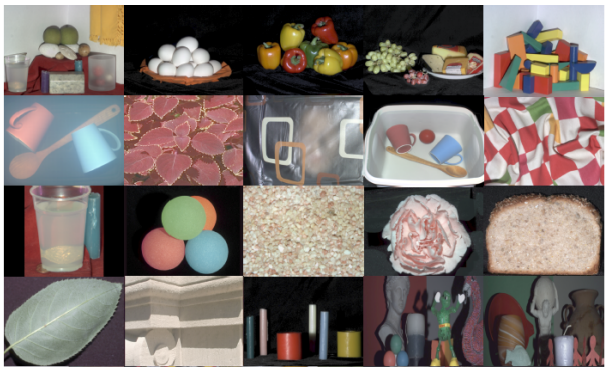

Figure 2: *Training images (direct and global components) gathered from prior work ($1-18$ from [NKGR06] and $19-20$ from [AN14]).*

4. Practical Considerations

Setup Equation 1 expresses the relation between direct and global reflectance in relation to the incident lighting $\mathbf{L}$, which are all expressed in the same parameterization (i.e., the projector's view in our case). Hence, the relation between each camera pixel and corresponding projector pixel location needs to be known. Our goal is to construct a setup that eases the calibration of the projection between camera and projector pixels. This projection is defined by the relative camera and projector's extrinsic and intrinsic parameters, as well as the scene geometry. To allow for direct-global separation on objects with ill-defined shape (e.g., fur), we desire a setup that does not require prior knowledge on the shape. Therefore, we opt for a co-axial camera-projector setup, where the transformation from projector pixel location to camera pixel location (and vice versa) is scene geometry independent, and it reduces to a simple scale that accounts for differences in resolution.

Our co-axial camera-projector setup consists of a $50-50$ beam splitter, a *Nikon D750* camera operating at 6016×4016 resolution, and an *Optoma W490* DLP projector operating at 1280×800 resolution. We follow a standard calibration procedure for the co-axial camera-projector pair. Please refer to the supplemental document for a detailed description of the geometric and radiometric calibration procedure.

Determining L As detailed in section 3 we synthesize our setup-specific training data from a small general set of publicly available decompositions and the target lighting $\mathbf{L}$. However, this lighting $\mathbf{L}$ is that received by the scene which includes any deviations from the ideal lighting introduced by the projector's optics *and* the beam splitter. Hence, we need to acquire the actual incident lighting $\mathbf{L}$.

To determine $\mathbf{L}$, we follow a data-driven approach and capture two photographs: one of a perpendicular (to the view direction) white diffuse plane $\mathbf{W_L}$ under the high frequency lighting pattern and another photograph $\mathbf{W}_1$ of the plane under uniform full-on

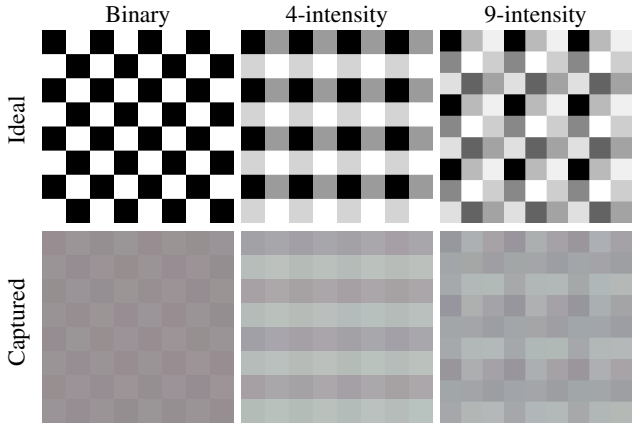

Figure 3: *Ideal versus projected lighting patterns. Due to limitations of off-the-shelf projectors, the actual contrast for high frequency lighting patterns is often reduced.*

lighting. The ratio $\mathbf{L}' = \frac{\mathbf{W_L}}{\mathbf{W_I}}$ eliminates any effects of imperfections and surface reflectance of the calibration surface. We then directly use this $\mathbf{L}'$ for generating the training data.

Lighting Pattern While our deep separation network can be trained for any lighting, not all will produce equally good results. As in prior work, we will rely on high frequency lighting to help disentangle the direct and global lighting. To maximize the granularity at which we can disentangle the lighting components, we desire to maximize the frequency content of the lighting patterns (i.e., at maximum projector resolution). Furthermore, we desire lighting patterns that are compatible with the deficiencies of off-the-shelf projectors (e.g., due to blur, light leakage, etc.).

To maximize the signal-to-noise ratio, we want to maximize the contrast difference between neighboring pixels. Ideally, a binary pattern fulfills this goal. However, due to optical limitations of off-the-shelf projectors, the ratio between "dark" and "light" pixels is limited (only 0.93 in our setup). To improve the average ratio, we consider regular 4-intensity and 9-intensity patterns, in repeated 2×2 or 3×3 pixel blocks respectively. However, due to the limitations of consumer-grade projectors, using a uniform distribution of pixel intensities does not necessarily yield the highest contrast. We therefore capture 2×2 (or 3×3) regular grid patterns with only one of the 4 (9 respectively) pixels turned on, and optimize the weights for each pixel, between 0 and 1, that yields the best ratio. Figure 3 shows the ideal and captured for the binary, 4-intensity, and 9-intensity patterns. For our set up, the average ratio between neighboring pixels for the 4-intensity pattern after optimization was 0.82, with a maximum of 0.93 and a minimum of 0.72. The 9-intensity pattern has a slightly lower average of 0.80, a similar lowest ratio but a much worse maximum ratio (0.99). We therefore opt to use the 4-intensity pattern.

Note all these calibration steps only need to be performed once when setting up the acquisition setup, and it can be reused for all acquisitions with the same setup and settings.

5. Results & Discussion

5.1. Qualitative Results

Figure 4 shows 8 decompositions of a variety of captured scenes at projector resolution. For comparison we also include decompositions obtained using the *multi-image* method of Nayar et al. [NKGR06]. However, we remark that this multi-image decomposition is *not* a ground truth decomposition; it is limited in accuracy by the frequency of the lighting patterns used. We briefly discuss each scene from top to bottom:

1. The *"Plastic Blocks"* scene includes a variety of plastic blocks in a transparent scale. Note that our method correctly decomposes the sticker on the orange block, and features the indirect lighting on the blue cylinder in the global component. A key difference is that our direct component is less colorful for a number of blocks. This is an example of the decomposition of subsurface scattering affected by the frequency of the lighting pattern; we will further analyze this in subsection 5.4.
2. The *"Candles & Soap"* example showcases the decomposition of a scene containing only translucent objects. Our results are qualitatively close to those obtained with Nayar et al.'s method.
3. The *"Ducky"* scene shows a stuffed animal next to a sharp specular cylinder and sphere. Here we can see a significant difference in the direct component on the stuffed duck due to the difference in lighting frequency. Another remarkable difference is that our deep separation method handles specular reflections better. The mirror image of the stuffed duck is very visible in Nayar et al.'s direct component, while in ours it is almost fully and correctly allocated to the global component.
4. The *"Glass Objects"* example shows two glass containers in front of a greeting card with one of the containers partially filled with water. Both decompositions look similar with exception of the caustic in the global component cast by the left container onto the top middle of the greeting card; the method of Nayar et al. exhibits the characteristic high frequency *"ringing"* artifacts. The method of Nayar et al. also produces an odd colorization of the decomposition of the water in the right container. However, our method loses a little bit of sharpness in the texture as it needs to aggregate information from neighboring pixels.
5. The *"Ball & Plants"* scene showcases complex interreflections between the ball and the planters, as well as subsurface scattering on the plants and pots. Again, we see high frequency light transport artifacts in Nayar et al.'s decomposition of the reflections of the orange ball on the planters in the global component, and differences in subsurface scattering decomposition of the planters (subsection 5.4).
6. The *"Fruit"* scene contains a variations of fruits. Our decomposition is qualitatively close to Nayar et al.'s decomposition, with some differences in the direct component most visible on the Star Fruit and Tangerine related to the decomposition of subsurface light transport (subsection 5.4).
7. The *"Colored Balls"* scene showcases the decomposition of strong indirect lighting between the three balls.
8. The *"Colored V-shapes"* shows diffuse interreflections between a series of v-shapes with colored sides.

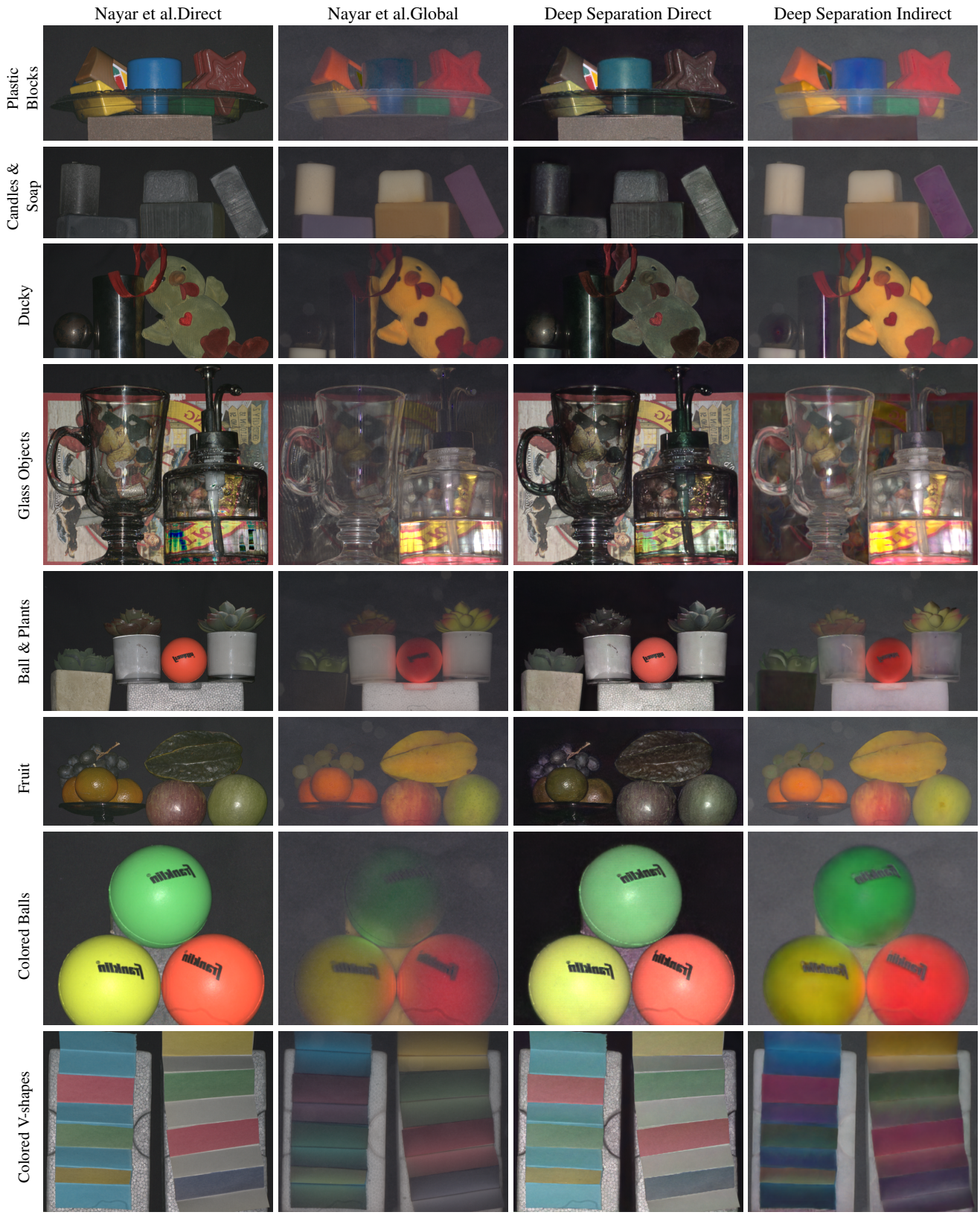

Figure 4: *A selection of scenes decomposed using the multi-image method of Nayar et al. [NKGR06] and our deep separation method at projector resolution.*

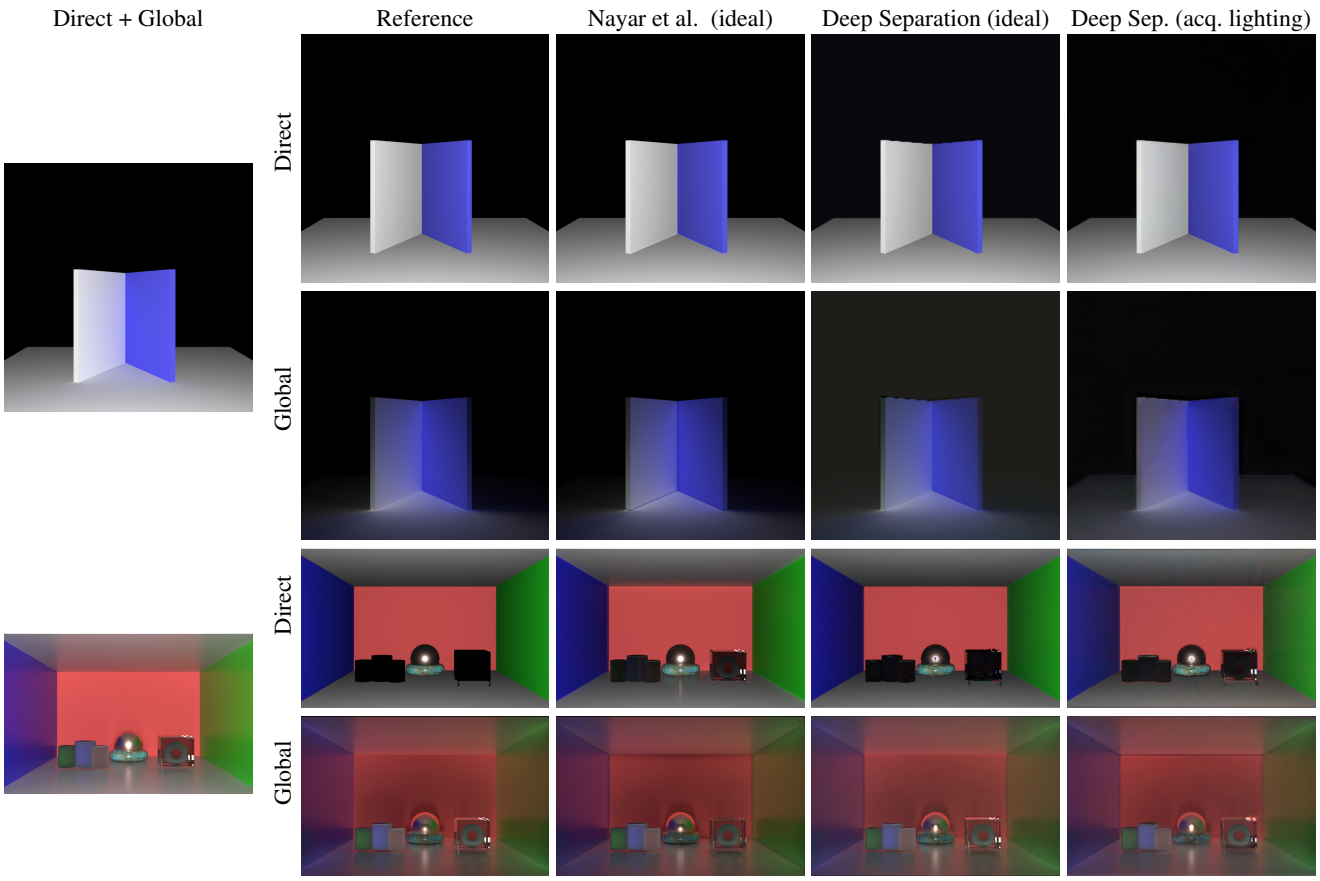

Figure 5: *Decompositions of two synthetic scenes (shown in the first column). Subsequent columns from left to right: ground truth reference decomposition, multi-image decomposition [NKGR06] with ideal lighting, deep separation with ideal lighting, and deep separation with captured lighting (i.e., including optical artifacts introduced by the projector).*

Table 1: *RMSE for the scenes shown in Figure 5.*

	V-Shape		Cornell Box	
	Direct	Global	Direct	Global
Ideal	0.0074	0.0174	0.1094	0.0419
Captured	0.0092	0.0106	0.1094	0.0504
[NKGR06]	0.0089	0.0081	0.0582	0.0571

5.2. Quantitative Results

The results in Figure 4 show that our method produces qualitatively good results. We further quantitatively corroborate the quality of our deep separation method on two synthetic scenes, featuring both glossy and diffuse interreflections. For each scene we synthesize a reference decomposition (i.e., global lighting is light transport that has undergone more than 1 bounce), a multi-image decomposition [NKGR06], a deep decomposition with the ideal lighting pattern, and a deep composition with a real-world captured lighting pattern (Figure 5). Note that the multi-image method has trouble correctly separating high frequency light transport (e.g., the apex of the v-shape, the glass object, and reflections on the glossy floor). Our deep separation method, on the other hand, loses some sharpness (e.g., on the textured torus). For each scene we also compute the RMSE (Table 1). From the error we can see that our method produces slightly higher errors than the multi-image method. In all, considering that our method decomposes from a single image, these errors are reasonable. Furthermore, we also observe that under ideal lighting our method performs slightly better than under the real-world degraded lighting pattern (as expected).

5.3. Comparison to Prior Single Image Methods

Figure 6 compares the quality of our results on the single image method proposed by Nayar et al. [NKGR06], the single image method of Subpa et al. [SaFZ*18], and the recent deep learning based method of Nie et al. [NGS*18]. For the single image method of Nayar et al., a single 4 pixel wide stripe pattern is used. We also follow Subpa et al. and use a 4×4 checkerboard pattern with a Fourier basis. For the deep learning method of Nie et al. we capture the scene under uniform white lighting. The single image methods of Nayar et al. and Subpa et al. suffer from the lower frequency of lighting pattern (e.g., reflection of the "Ducky" in the direct component) as well as interpolation artifacts (i.e., blocky and ringing respectively). The method of Nie et al., while plausible, does not produce an accurate decomposition. However, it should be noted

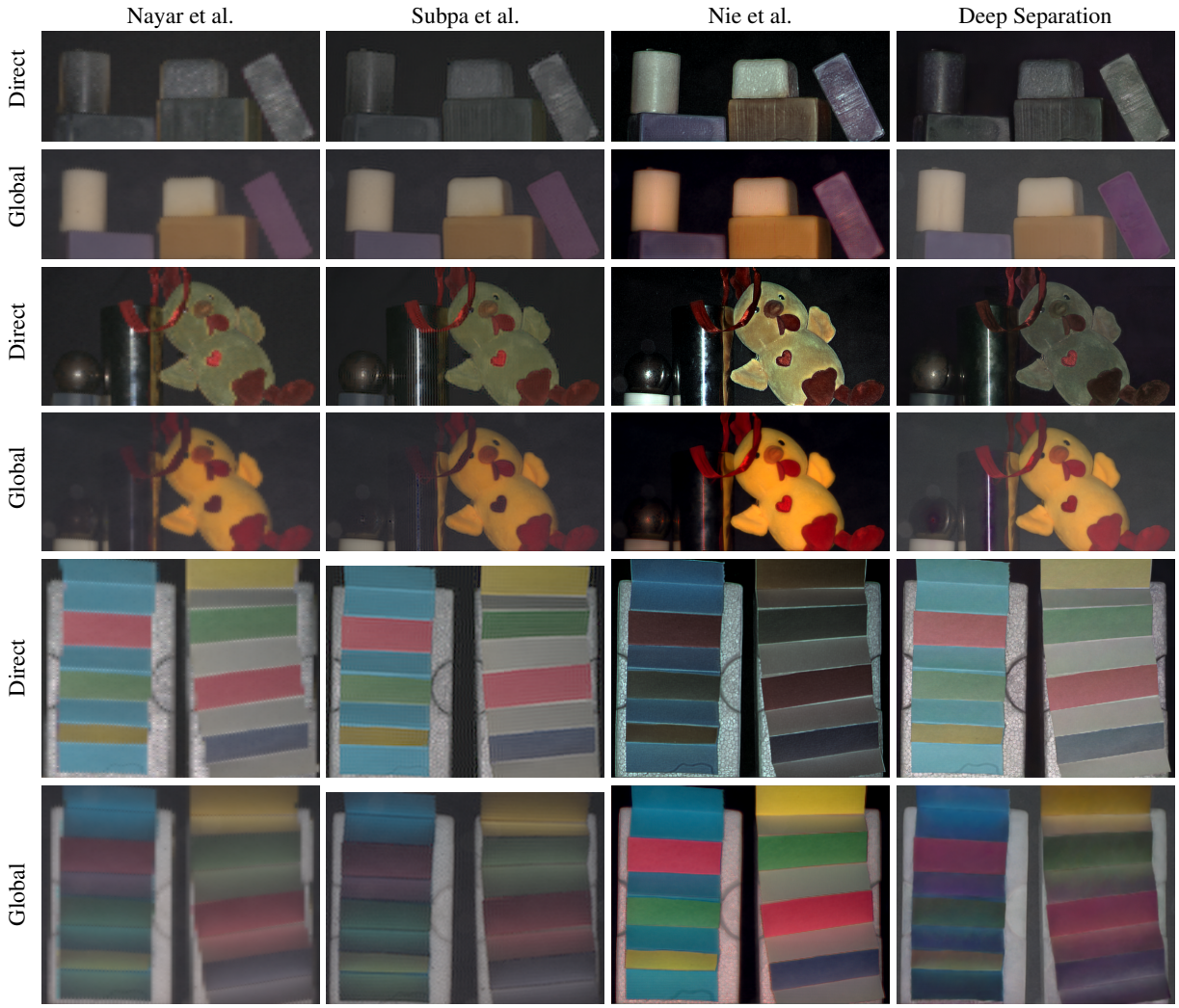

Figure 6: *Comparison of our deep decomposition results against the single image decomposition methods of Nayar et al. [NKGR06], Subpa et al. [SaFZ*18], and Nie et al. [NGS*18].*

that the method of Nie et al. solves a much more difficult problem, namely a decomposition under unstructured lighting. We can also train our network under uniform full-on white lighting, but this does not yield as good results as Nie et al. (please refer to the supplemental document for an example).

5.4. Impact of Lighting Frequency

As noted in subsection 5.1, for scenes with subsurface scattering as well as scenes that exhibit high frequency light transport our method produces different results that the multi-image method of Nayar et al. The main reason for this difference is the higher frequency of our structured lighting. For the multi-image decomposition we used 4-pixel wide stripe patterns, effectively illuminating the scene with a 4 times lower frequency pattern. Note that using a sinusoidal lighting pattern for the method of Nayar et al. does not alleviate this issue since the frequency of the sinusoids is still band limited; it just produces slightly smoother artifacts for high frequency light transport. While a better decomposition of high frequency light transport is expected when increasing the frequency of the lighting patterns, its impact on subsurface scattering is perhaps less obvious. Ideally, all subsurface transport should be classified as global transport, and the main source of direct lighting is direct (white) reflections due to the dielectric surface reflectance. In practice, the frequency of the lighting pattern essentially determines which scattered reflectance is consider direct "single scattering" and which ones is global "multiple scattering". If the pattern is low frequency, more reflectance will be allocated to the direct component; in the limit case for very low frequent lighting, the minimum observed value will be black and thus no indirect lighting is "observed" and the decomposition assigns all subsurface scattering to the direct component. To further illustrate this, we also captured our scenes with a 4 times lower resolution lighting pattern. Because we decompose our image at "projector" resolution (4 times lower in

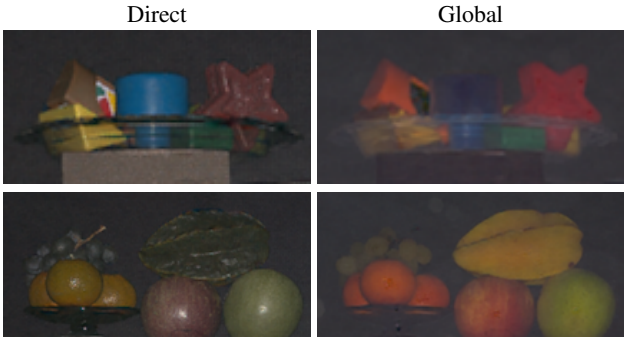

Figure 7: *Deep separation with 4 times lower frequency lighting. The resulting decomposition is more similar (in color in the direct component) to the multi-image decomposition as shown Figure 4.*

this case), the resulting decomposition are also of lower resolution. Albeit at lower resolution, Figure 7 illustrates that the decompositions, and in particular the color of the direct component of the subsurface scattering materials, is more similar to the multi-image method of Nayar et al. [NKGR06] in Figure 4.

5.5. Ablation Study

Generalization Typically, training a neural network requires a large set of training exemplars to avoid overfitting and ensure generalization to unseen cases. Yet, we only rely on just 20 decompositions as training exemplars, albeit with an extensive data-augmentation step. We perform three ablation experiments to show that our method generalizes well to unseen cases.

Our first indication that our method generalizes well is that we only train on captured decompositions, yet, as shown in Figure 5 our method works without any adaptation, beyond the calibration steps, on synthetic scenes. These synthetic scenes are significantly different than the training data.

Second, to further demonstrate the necessity of our data-augmentation, we also trained a network without applying color swapping and color scaling augmentation. Figure 8 shows that without swapping, severe artifacts are visible. While less visible, without color scaling, the resulting network fails to adequately decompose the glossy indirect reflections on left and right walls of the synthetic scene.

As a final experiment, we train a network without using any of the training exemplars that exhibit subsurface scattering. For the scenes in Figure 2, we remove any training patches that overlap with the candles in scenes 1, 11, 19, and 20, and we completely remove all training data from scenes 3, 4, 12, and 18. Despite removing a significant portion of the training data, we observe in Figure 9 that the resulting network is still able to correctly decompose subsurface light transport.

Loss Function Thanks to our extensive data-augmentation we can rely on a straightforward direct difference loss between the reference and predicted solutions to train our decomposition network. Perhaps less obvious is the decision to compute the distance in LAB

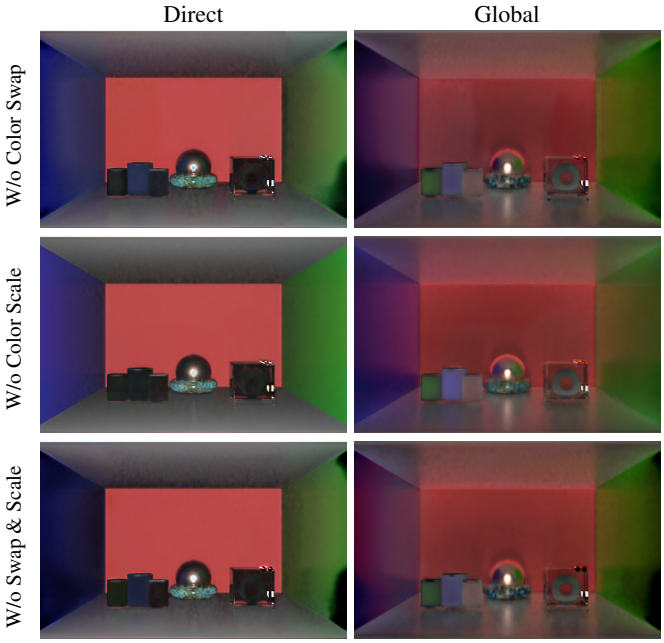

Figure 8: *Color channel swapping and color scaling training exemplar augmentation is essential for generalization to unseen scenes (i.e., compared to Figure 5).*

Figure 9: *To illustrate the generalization capabilities of our method, we decompose a scene with subsurface scattering with a network trained* without *subsurface scattering exemplars.*

space as opposed to RGB, and to use the square distance (L_2) instead of the L_1 distance. Empirically, we found that both the L_1 and L_2 distance in RGB can produce artifacts at sharp high contrast edges, and that an L_1 LAB loss produced artifacts in the global decomposition of smoothly varying regions. We refer to the supplemental material for a visual comparison.

Joint Inference of Direct and Global Our network uses a shared encoder and features two decoders. This raises the question whether we need to output both components, or whether it is possible to only compute a single component and then leverage Equation 1 to compute the other component. We therefore, train a network that only outputs the direct component $\mathbf{I_d}$, and compute the global component as: $\mathbf{I_g} = \frac{1}{w(\mathbf{L})}(\mathbf{I} - \mathbf{I_d}\mathbf{L})$. As can be seen in Figure 10, this produces a global component where the lighting pattern is still visible, because Equation 1 is only approximate.

Cross Connections Our network features cross connections to share information between the layers of the direct and global de-

Table 2: *RMSE for the scenes shown in Figure 5 for decomposition networks trained for different training patch sizes and layer configurations.*

	V-shape				Cornell Box			
	Repeated Layer		Single Layer		Repeated Layer		Single Layer	
	Direct	Global	Direct	Global	Direct	Global	Direct	Global
128×128	0.0092	0.0106	0.0117	0.0104	0.1094	0.0504	0.1085	0.0561
64×64	0.0095	0.0130	0.0185	0.0281	0.1089	0.0547	0.1107	0.0578
32×32	0.0208	0.0231	0.0178	0.0203	0.1069	0.0519	0.1077	0.0498
16×16	0.0206	0.0290	0.0257	0.0220	0.1076	0.0516	0.1207	0.0563

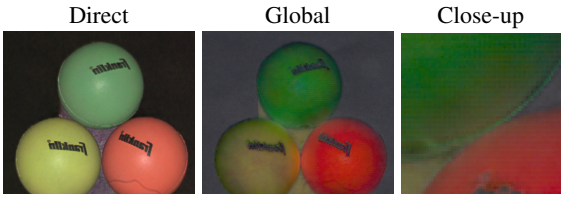

Figure 10: *Only estimating the direct component with a deep network, and then estimating the global component using Equation 1 does not completely remove the structured lighting pattern from the global component.*

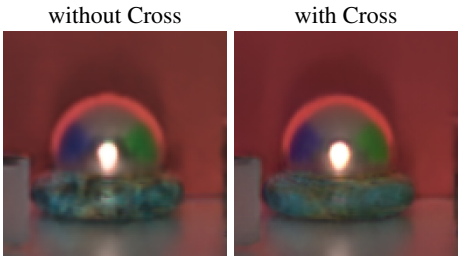

Figure 11: *Impact of cross connections between the direct and global decoder networks.*

coder networks. While the differences are relatively subtle, we found that it produces cleaner decompositions for high frequency textures. Figure 11 illustrates that without cross connections the high frequency texture on the torus is not as well decomposed.

Training Patch Size The depth of our network depends on the training patch size (128×128 in our implementation), and the number of repeated layers per resolution (2 convolution layers in our implementation). To better understand the impact of both factors, we perform an ablation study on the synthetic scenes where we reduce the training patch size (from 128 to 16) and with a single or two convolutional layers per resolution. Table 2 summarizes the RMSE for the synthetic *Cornell Box* and *V-shape* scenes. We refer to the supplemental material for a visual comparison on the *Cornell Box* scene for the different network configurations. From Table 2 we can see that our selected configuration (128×128 with two convolutional layers) performs best. However, the single convolutional layer at 128×128 and both configurations at 64×64 are also competitive, and can serve as an alternative that require less training time. The 32×32 solutions can also produce high quality results, albeit it not equally over all scenes. At 16×16 the RMSE error significantly increases, and visually we can observe discolorations in the decompositions.

6. Resolution Augmentation

Our deep learning based solution currently assumes that the camera image is expressed at projector resolution. Typically, however, cameras have a much high resolution than projectors. It is therefore desirable to compute the decomposition at camera resolution rather than projector resolution.

We observe that our network architecture presented in section 3 does not explicitly assume that the lighting $\mathbf{L}$ and capture photograph $\mathbf{I}$ are expressed at projector resolution. Indeed we can also train the network with both $\mathbf{L}$ and $\mathbf{I}$ expressed at camera resolution. While this provides a reasonable separation (Figure 12 left), the lighting pattern is still visible in the decomposition. A common strategy to reduce the effect of such high frequency distortions is to add to the loss function a total variation penalty term: $|||\nabla \mathbf{I_d}| + |\nabla \mathbf{I_g}|||_1$. However, such a term tends to also remove the high frequency features from the direct and global components (Figure 12 middle). Instead, we only want to remove the high frequency features introduced by the lighting pattern. We therefore only apply a total variation loss when the lighting pattern introduces a high frequency change:

$$E_{full} = E_{sep} + w|||\nabla \mathbf{L}|(|\nabla \mathbf{I_d}| + |\nabla \mathbf{I_g}|)||_1, \quad (3)$$

where the L_1 norm sums over all camera pixel, and w is an appropriate weighting factor. In our implementation we set w to 0.00004. We train on 128×128 (camera resolution) patches, which effectively results in a "smaller" patch in projector resolution. Modulating gradients of the direct and global component with the gradients of the lighting produces sharper decompositions (Figure 12 right).

Figure 13 shows for a selection of scenes from Figure 4 a full camera resolution decomposition. For comparison, we also show a projector-resolution decomposition that has subsequently been upsampled using a state-of-the-art deep super-resolution network [ZLDQ19]. As can be seen, while our method exhibits slightly less detail than the multi-image decomposition, it features significantly more detail (and less artifacts) than the super-resolution result. Furthermore, our method can, thanks to the full projector-resolution lighting patterns, more accurately decompose high frequency light transport paths.

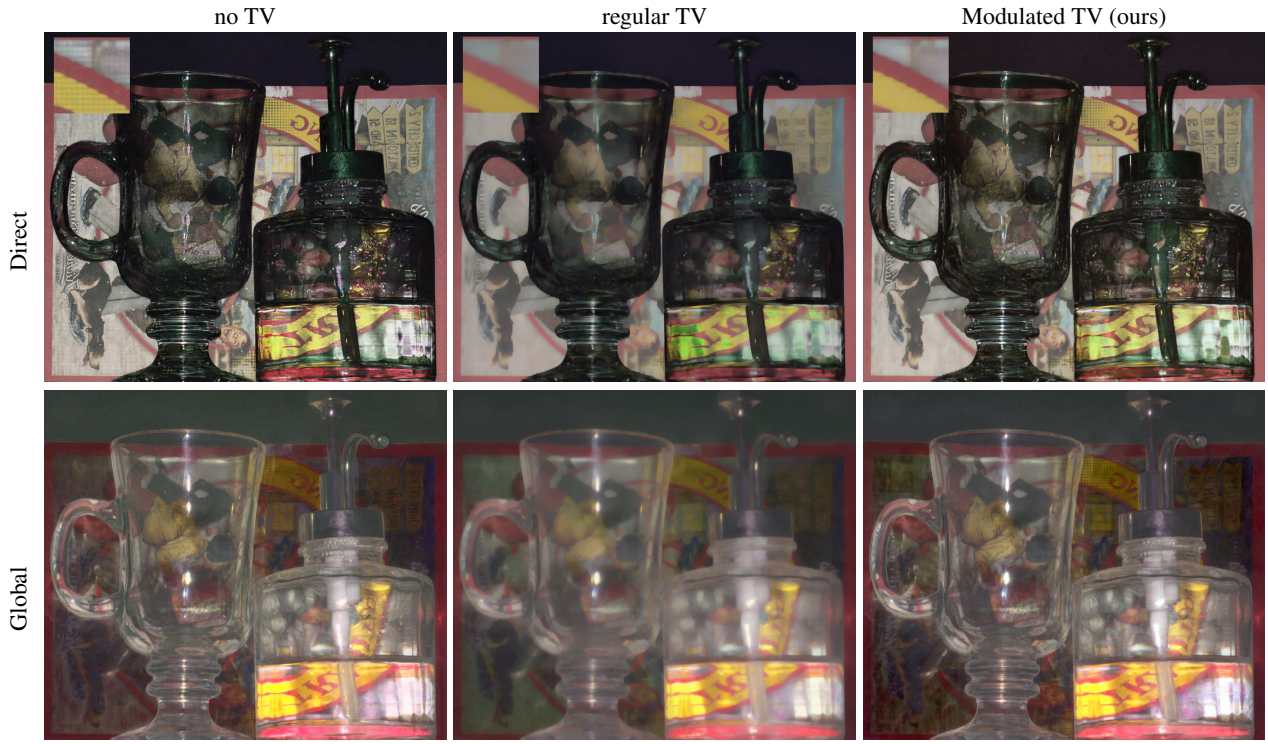

Figure 12: *The deep separation network directly trained on full camera resolution fails to fully remove the structured lighting pattern. Adding a regular total variation loss removes the structured lighting pattern, but also over-smooths. Our modulated total variation strikes a balance between smoothing and removing the effects of the structured lighting.*

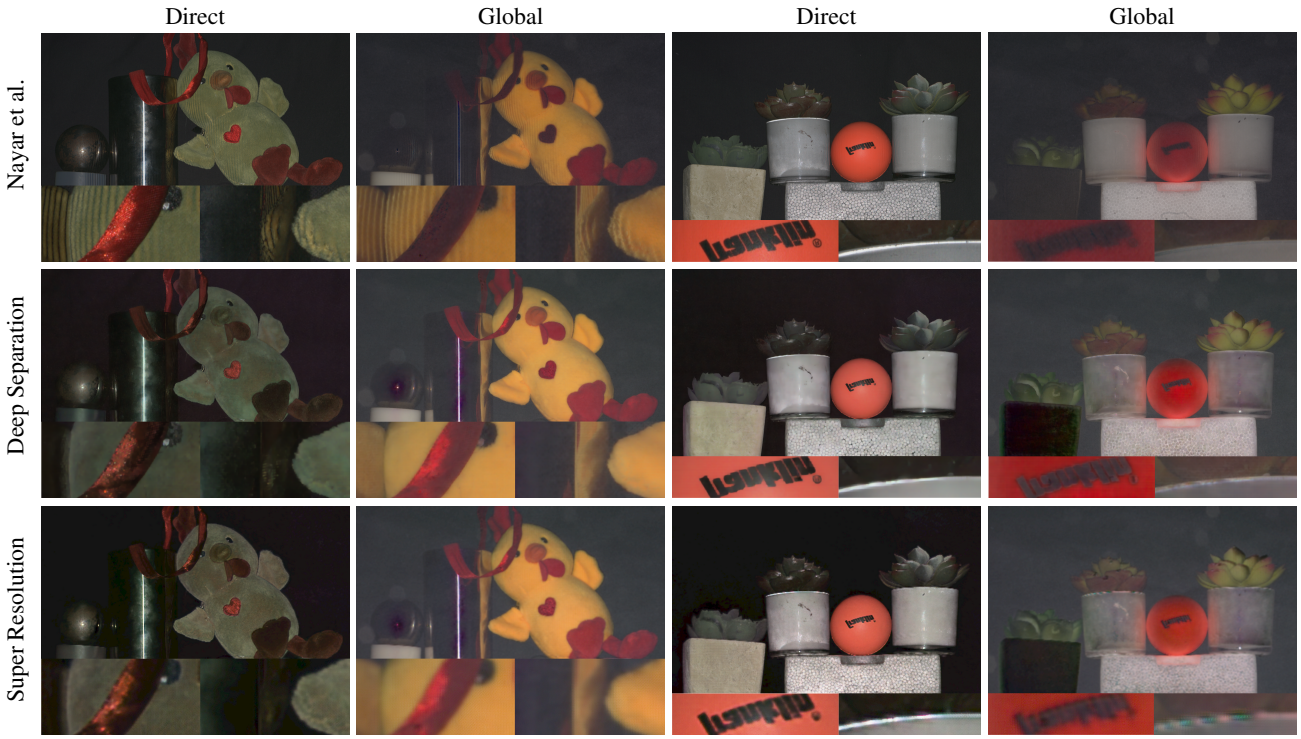

Figure 13: *Comparison between full camera resolution decompositions using a multi-image method, our deep separation method, and a super-resolution method [ZLDQ19] applied to a deep projector-resolution decomposition.*

7. Limitations

Our method is not without limitations. Two of our main limitations are related to our setup. First, the quality of camera image plays a significant role. A high quality (i.e., without compression artifacts and minimal noise) estimate of pixels values is required to produce good results. Second, off-the-shelf projectors are designed to project on a planar projection screen, and therefore typically have a limited depth of field. This also limits the depth of the scene for which we can faithfully decompose the direct and global components. this can potentially be overcome by either using a projector with a larger depth of field (e.g., a laser-projector), or potentially by using multiple networks trained for decomposing the photograph at a specific depth, and then combining the images with a second network. An interesting avenue for future research would be to relax the co-location requirement (i.e., beam-splitter setup) so that the difference between depths is increased (cf. disparity), and thus aiding in combining the decompositions with a second network.

Another limitation is that our decompositions exhibit some spatial degradation and some loss of quality because the high frequency lighting introduces spatial variations that the network needs to correct.

Finally, while our method can attain real-time decompositions (e.g., on a NVidia Quadro P6000, a 500×200 image can be decomposed at 50 fps, and a 4096×1024 image at 20 fps), our method is not explicitly trained to produce temporally coherent decompositions on video sequences. In the supplemental material we demonstrate a decomposition of a synthetic scene with a diffuse ball rolling from one side to the other. Despite not being explicitly trained to be temporally coherent, we can see that the decomposition is mostly coherent. The global component exhibits some flickering, especially on the rolling ball. Note that our method correctly and coherently decomposes the glossy reflections on the green and blue wall, as well as the indirect shadows on the background.

8. Conclusion

In this paper we presented a deep learning based solution for decomposing the direct and global light transport components of a scene from a single photograph captured under high frequency illumination. Our method is designed to operate on non-binary lighting patterns, allowing us to naturally compensate for the deficiencies in common off-the-shelf projectors. Furthermore, this also allows us to separate the direct and global components at full projector resolution, yielding more accurate decompositions for lighting frequency sensitive features such as subsurface scattering and specular indirect lighting.

Avenues for future research include adapting our method to operate on non-colocated setups, to overcome the depth of field limitations of common projector systems, to improve temporal coherency, and to improve robustness by using multiple lighting patterns.

Acknowledgments We thank the reviewers for their constructive feedback and Nie et al. [NGS*18] for providing us with direct-global decompositions (Figure 6) using their single image network solution. This work was partially supported by NSF grant IIS-1909028.

References

[AN14] ACHAR S., NARASIMHAN S. G.: Multi focus structured light for recovering scene shape and global illumination. In *ECCV* (2014), pp. 205–219. 2, 4

[ANN13] ACHAR S., NUSKE S. T., NARASIMHAN S. G.: Compensating for motion during direct-global separation. In *ICCV* (2013), pp. 1481–1488. 2

[GKGN11] GU J., KOBAYASHI T., GUPTA M., NAYAR S. K.: Multiplexed illumination for scene recovery in the presence of global illumination. In *ICCV* (2011), pp. 691–698. 2

[GTNZ12] GUPTA M., TIAN Y., NARASIMHAN S. G., ZHANG L.: A combined theory of defocused illumination and global light transport. *IJCV 98*, 2 (2012), 146–167. 2

[KB14] KINGMA D. P., BA J.: Adam: A method for stochastic optimization. *ICLR abs/1412.6980* (2014). 4

[KYN12] KOPPAL S. J., YAMAZAKI S., NARASIMHAN S. G.: Exploiting DLP illumination dithering for reconstruction and photography of high-speed scenes. *IJCV 96*, 1 (2012), 125–144. 2

[NGS*18] NIE S., GU L., SUBPA-ASA A., KACHER I., NISHINO K., SATO I.: A data-driven approach for direct and global component separation from a single image. In *ACCV* (2018), pp. 133–148. 3, 7, 8, 12

[NKGR06] NAYAR S. K., KRISHNAN G., GROSSBERG M. D., RASKAR R.: Fast separation of direct and global components of a scene using high frequency illumination. *Trans. Graph. 25*, 3 (2006), 935–944. 1, 2, 3, 4, 5, 6, 7, 8, 9

[OANK15] O'TOOLE M., ACHAR S., NARASIMHAN S. G., KUTULAKOS K. N.: Homogeneous codes for energy-efficient illumination and imaging. *ACM Transactions on Graphics (ToG) 34*, 4 (2015), 35. 2

[OHX*14] O'TOOLE M., HEIDE F., XIAO L., HULLIN M. B., HEIDRICH W., KUTULAKOS K. N.: Temporal frequency probing for 5d transient analysis of global light transport. *Trans. Graph. 33*, 4 (2014), 87. 2

[OMK16] O'TOOLE M., MATHER J., KUTULAKOS K. N.: 3d shape and indirect appearance by structured light transport. *IEEE PAMI 38*, 7 (2016), 1298–1312. 1, 2

[ORK12] O'TOOLE M., RASKAR R., KUTULAKOS K. N.: Primal-dual coding to probe light transport. *Trans. Graph. 31*, 4 (July 2012). 1, 2

[RRC12] REDDY D., RAMAMOORTHI R., CURLESS B.: Frequency-space decomposition and acquisition of light transport under spatially varying illumination. In *ECCV* (2012), pp. 596–610. 2

[SaFZ*18] SUBPA-ASA A., FU Y., ZHENG Y., AMANO T., SATO I.: Separating the direct and global components of a single image. *Jour. Inf. Proc. 26* (2018), 755–767. 1, 2, 7, 8

[WVO*14] WU D., VELTEN A., O'TOOLE M., MASIA B., AGRAWAL A., DAI Q., RASKAR R.: Decomposing global light transport using time of flight imaging. *IJCV 107*, 2 (2014), 123–138. 2

[ZLDQ19] ZHANG W., LIU Y., DONG C., QIAO Y.: Ranksrgan: Generative adversarial networks with ranker for image super-resolution. In *ICCV* (2019), pp. 3096–3105. 10, 11

[ZPIE17] ZHU J.-Y., PARK T., ISOLA P., EFROS A. A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV* (2017). 3