Reflection Removal Using a Dual-Pixel Sensor

Abhijith Punnappurath
York University
pabhijith@eecs.yorku.ca

Michael S. Brown
York University
mbrown@eecs.yorku.ca

Abstract

Reflection removal is the challenging problem of removing unwanted reflections that occur when imaging a scene that is behind a pane of glass. In this paper, we show that most cameras have an overlooked mechanism that can greatly simplify this task. Specifically, modern DLSR and smartphone cameras use dual pixel (DP) sensors that have two photodiodes per pixel to provide two sub-aperture views of the scene from a single captured image. "Defocus-disparity" cues, which are natural by-products of the DP sensor encoded within these two sub-aperture views, can be used to distinguish between image gradients belonging to the in-focus background and those caused by reflection interference. This gradient information can then be incorporated into an optimization framework to recover the background layer with higher accuracy than currently possible from the single captured image. As part of this work, we provide the first image dataset for reflection removal consisting of the sub-aperture views from the DP sensor.

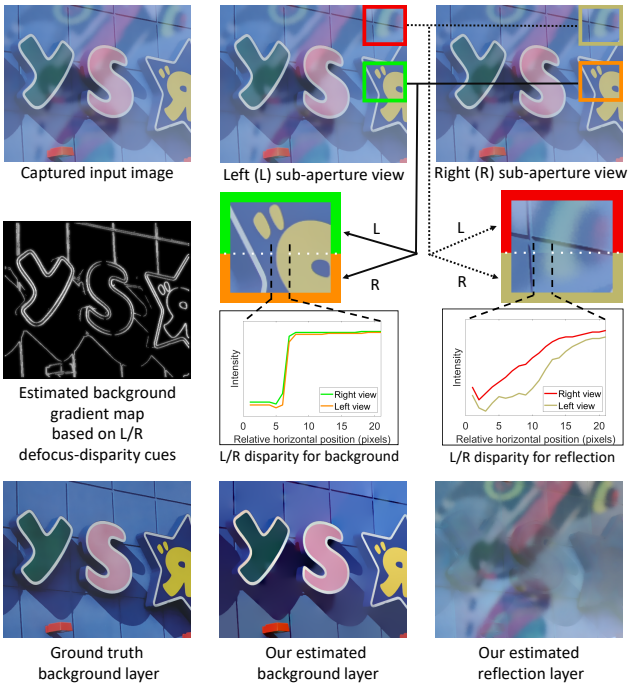


Figure 1. An example sketching our basic idea. The captured image and its two sub-aperture views are shown. In the zoomed-in boxes, the upper half corresponds to the left view, and the lower half to the right. In the box on the right showing an out-of-focus reflection region, a horizontal shift can be observed between the two white dots (best viewed electronically and zoomed), while no disparity exists in the left box of an in-focus background region. This disparity (illustrated in the plots) allows us to compute a mask for image gradients belonging to the background region that can be used to extract the background layer.

1. Introduction

This paper addresses the problem of removing reflection interference that occurs when imaging a scene behind a pane of glass. The novelty of our work lies in our use of the information available from dual pixel (DP) sensors that are found on most smartphone and DSLR cameras. Traditional image sensors have a single photodiode per pixel site. DP sensors have two photodiodes that effectively split the pixel in half. The DP sensor design furnishes, from a *single* captured image, *two* views of the scene where rays passing through the left side of the lens are captured by the right half-pixels (right sub-aperture view) and those passing through the right side of the lens are captured by the left half-pixels (left sub-aperture view).

The DP sensor is effectively a rudimentary two-sample light-field camera. Within this context, scene points that are in-focus will have no difference between their positions in the left and right sub-aperture views. However, out-of-focus scene points will be blurred in opposite directions in the two sub-aperture views, resulting in very small but detectable shifts. These shifts, which we refer to as defocus-disparity cues, are related to the amount of out-of-focus blur incurred by the scene point with respect to the camera lens's depth of field. These defocus-disparity cues, which are natural by-products of the DP sensor, allow us to robustly determine which gradients in the captured composite image belong to the in-focus background layer. Fig. 1 shows an example.

Contribution We introduce a new reflection removal method that exploits the two sub-aperture views available on a DP sensor. We explain the relationship between the defocus-disparity cues in the two sub-aperture views with respect to the background layer and the objects reflected by the glass. Working from this backdrop, we propose a method that uses these defocus-disparity cues to detect gradients corresponding to the in-focus background and incorporate them into an optimization framework to recover the background layer. Our experimental results demonstrate the advantages of this additional information over current methods. More importantly, our results are obtained without hardware modifications or training – we simply use the data that was already available, yet ignored. As part of this work, we introduce a new dataset for reflection removal that provides access to the two sub-aperture views.

1.1. Related work

We first provide a brief overview of the original functionality of DP sensors, and their extended capabilities. We also discuss single- and multi-image reflection removal methods as well as methods using light-field cameras as the DP sensor can be considered a two-sample per pixel light-field.

DP sensor Dual pixel sensors were developed to provide a fast method for autofocus [13, 28], the idea here being that by examining the image disparity between the two views, a change in lens position can be calculated to minimize the amount of out-of-focus blur, thus focusing the image. However, the DP data can be used for tasks beyond autofocus. Recent work by Wadhwa et al. [30] showed how the two DP views can be used to extract dense depth maps for the purpose of synthesizing shallow depth-of-field images.

Reflection removal

Single image Most single-image methods exploit the statistics of natural images to make the reflection removal problem less ill-posed. Long-tail distribution of gradients [18], sparsity of corners and edges [20], the ghosting effect [27], difference in smoothness between the background and reflection layers [22], and depth of field confidence maps [33] are some of the priors that have been employed.

More recently, deep learning techniques have also been applied to this task [9, 40, 32, 39]. Fan et al. [9] first learn an intermediate edge map to guide background recovery, whereas Wan et al. [32] combine the two stages of gradient and image inference into a unified framework. While Zhang et al. [40] seek to use both low- and high-level image information, Yang et al. [39] estimate both the background and the reflection layers in cascade. Although much progress has been made in single-image reflection removal, there is still a large margin for improvement due to the highly ill-posed nature of the problem.

Multiple images Capturing multiple images of the scene in a pre-defined manner can make the reflection removal problem more tractable. The vast majority of multi-image methods are based on motion cues [12, 5, 10, 11, 21, 29, 37]. These methods take advantage of the difference in motion between the two layers given images of the same scene taken from different viewpoints. Prior works have modeled the motion of the two layers as pure translation [5], affine [10], or a full homography [11]. Recent approaches [12, 21, 29, 37] have replaced these parametric models with dense per-pixel motion vectors. Methods that require specialized hardware or non-conventional capture settings have also been proposed – using a polarizer [26, 24, 15, 8], varying focus [25], capturing a flash no-flash pair [2, 3] and so forth. Although these multi-image approaches produce better results due to the availability of additional information, they place the burden on the photographer to acquire special hardware or skills, and thereby vastly limit their applicability to lay users.

Light-field cameras While layer separation is ill-posed with conventional imaging, the task becomes tractable with light field imaging as demonstrated by recent works [34, 14, 6, 23]. Wang et al. [34] built their own portable camera array to obtain an image stack for reflection removal. Johannsen et al. [14] propose a variational approach for layer separation assuming user assistance in gradient labeling. Chandramouli et al. [6] advocate a deep learning approach to recover the scene depth, as well as the two layers. Ni et al. [23] use focus manipulation to remove the reflections. The drawback of these methods is the need for specialized light field cameras. Our method, in contrast, works by using information available on DP sensors of most current commodity cameras.

2. Image formation model with DP sensor

A DP sensor splits a single pixel in half using an arrangement of a microlens sitting atop two photodiodes. See Fig. 2(a). The two halves of the dual pixel – the two photodiodes – can detect light and record signals independently. When the two signals are summed together, the pixel intensity that is produced will match the value from a normal single diode sensor. The split-pixel arrangement has the effect that light rays from the left half of the camera lens's aperture will be recorded by the right half of the dual pixel, and vice versa.

A scene point that is out of focus will experience a disparity or shift between the left and right views due to the circle of confusion that is induced. It is precisely this shift that is exploited by DP auto-focus systems. By examining the *signed* average disparity value within a region of interest, the auto-focus algorithm can determine not only in which direction to move the lens in order to bring that region into focus (and thereby minimize disparity) but also by how much.

Within this backdrop, we examine the image formation model for a DP sensor imaging a scene through a trans-

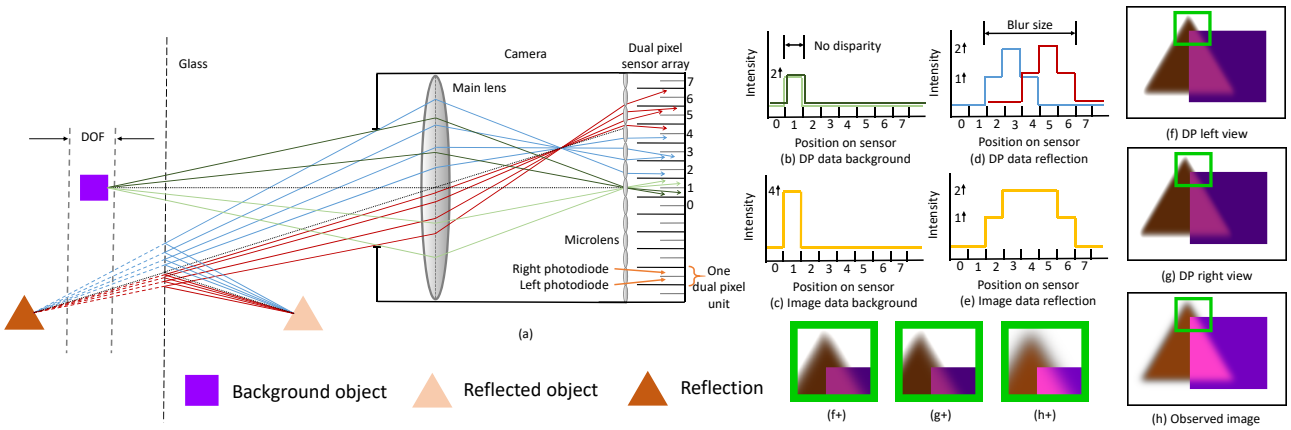


Figure 2. Image formation model for a dual-pixel camera capturing a scene behind glass. (a) An in-focus background scene point is recorded at pixel site 1, while an out-of-focus reflection scene point creates a defocus blur spread across pixels sites 2 to 6. Light from opposite halves of the lens is collected by the left and right half-pixels. There is no disparity for the background scene point (b), whereas a disparity proportional to the blur size is induced by the reflection (d). The sum of the left and right half-pixels represents the observed image intensity at that pixel site (c), (e). The DP view and the observed image are a superposition of the background object and the reflected object (f-h). The shift in the reflection between the two views is evident from the position of the tip of the triangle (f+,g+).

parent glass. A dense DP sensor array effectively yields views through the left and right halves of the lens from a single capture. Depending on the sensor's orientation, this can also be the upper and lower halves of the lens; without loss of generality, we consider them to be the left and right views in the rest of the paper.

We make the following two assumptions. First, we assume the background layer has predominately stronger image intensity than the reflection layer. This assumption is made by all reflection removal algorithms. Second, we assume the background scene content lies within the depth of field (DOF) of the camera, while the objects in the scene being reflected on the glass are at a different depth and therefore outside the DOF. The second assumption is also common [22, 33, 38, 7, 9, 39], and as noted by Wan et al. [31], it is quite reasonable to presume that the background and the objects in front of the glass have different distances from the camera. In such a scenario, the observed image is a superposition of the in-focus background and a de-focused reflection.

Based on these assumptions, we illustrate the image formation model in Fig. 2(a) for a DP camera imaging a scene through glass. A point on the in-focus background object emits light that travels through the camera's lens and is focused on the sensor at a single pixel (labeled as 1). Observe from the figure that rays that pass through the right half of the main lens aperture hit the microlens at an angle such that they are directed into the left half-pixels. The same applies to the left half of the aperture and the right half-pixels. For an in-focus scene point, there is no disparity (Fig. 2(b)). The sum of the left and right values is stored as the image intensity at that pixel (Fig. 2(c)).

Next, consider the triangular object in front of the glass that constitutes the reflection layer. Light from a point on this object focuses in front of the sensor, and produces a five-pixel wide (labeled 2 to 6) *defocus-blurred* image on the sensor. The left and right views created by the split-pixels have a disparity that is proportional to the blur size (Fig. 2(d)). The blurred reflection image is obtained by summing up the left and right signals (Fig. 2(e)). The composite DP data that is the sum of the in-focus background (with no disparity) and the out-of-focus reflection (with a non-zero disparity) as observed from the left and right views over the entire imaging sensor is shown in Figs. 2(f,g). Notice the shift between views as highlighted by the zoomed-in regions (f+,g+). The final image output by the camera that is the sum of left and right DP views is also shown in Fig. 2(h), and its zoomed-in region in (h+).

If $\mathbf{b}$ represents the background layer and $\mathbf{f}$ denotes the latent sharp reflection layer, both in lexicographically ordered vector form, the composite left $\mathbf{g}_{\text{LV}}$ and right $\mathbf{g}_{\text{RV}}$ DP views can be expressed mathematically as

$$\mathbf{g}_{\text{LV}} = \frac{\mathbf{b}}{2} + \mathbf{W}_{\text{LV}}\mathbf{f}, \;\; \mathbf{g}_{\text{RV}} = \frac{\mathbf{b}}{2} + \mathbf{W}_{\text{RV}}\mathbf{f}, \tag{1}$$

where $\mathbf{W}_{\text{LV}}$ and $\mathbf{W}_{\text{RV}}$ are the matrices that multiply the underlying sharp reflection layer $\mathbf{f}$ to produce its defocused and shifted versions of half intensity in the left and right views, respectively. The observed image $\mathbf{g}$ can be expressed as $\mathbf{g} = \mathbf{g}_{\text{LV}} + \mathbf{g}_{\text{RV}} = \mathbf{b} + \mathbf{r}$, where $\mathbf{r}$ equals the blurred reflection layer and is given by $\mathbf{r} = (\mathbf{W}_{\text{LV}} + \mathbf{W}_{\text{RV}})\mathbf{f}$.

Figure 3. Input images and our estimated weighted gradient maps of the background.

3. Proposed method

Working from our previous section, we describe our reflection removal method that leverages the *a priori* knowledge that (i) the background layer is sharp and has zero disparity, and (ii) the reflection layer is defocus-blurred and has a non-zero disparity between the left and right DP views.

3.1. Defocus-disparity cues

Levin et al. [18] demonstrated that labeling the gradients of the input image can serve as a powerful mechanism for reflection removal. However, the labeling was done manually by the user. Inspired by the success of [18], we propose to use the defocus-disparity cues between the left and right DP views to automatically identify which gradients belong to the background layer.

Let the gradients of the left and right DP views obtained by applying the first-order horizontal and vertical derivative filters be represented as $\mathbf{h}_{\text{LV}}$ and $\mathbf{h}_{\text{RV}}$. To compute disparity, we select a patch of size $N \times N$ pixels in $\mathbf{h}_{\text{LV}}$ and perform a horizontal search over a range of $-t$ to t pixels in $\mathbf{h}_{\text{RV}}$. A 1D search suffices because the split-pixels produce an almost pure horizontally rectified disparity in the sensor's reference frame. The search interval $2t+1$ can be restricted to a few pixels because the baseline between DP views is very narrow (approximately equal to aperture diameter [30]). We compute the sum of squared differences (SSD) for each integer shift. Following [30], we find the minimum of these $2t+1$ points and fit a quadratic $\frac{1}{2}a_1x^2+a_2x+a_3$ to the SSD value using the minimum and its two surrounding points. At a given pixel i, the location of the quadratic's minimum $s_i = \frac{-a_2}{a_1}$ serves as our sub-pixel minimum. We also compute a confidence value at each pixel i as [4]:

$$\beta_i = \exp\left(\frac{\log|a_{1_i}|}{\sigma_{a_1}} - \frac{a_{3_i}}{\sigma_{a_3}^2}\right). \tag{2}$$

We construct our weighted gradient map of the background using the confidence values β_i as

$$c_i = \begin{cases} \rho\beta_i & \text{if } |s_i| < \epsilon \text{ and } \beta_i > 1, \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

Two examples of our estimated background gradient maps are shown in Fig. 3. We fix $\rho = 5$, $N = 11$, $t = 5$, $\sigma_{a_1} = 5$, and $\sigma_{a_3} = 256$ for all examples presented in this paper. Note that blurred reflection gradients are weak [22, 38, 7], and very few can be reliably labeled. In our experiments, we did not observe any improvement in the results by adding labeled reflection gradients to our cost function, and therefore, we do not include them in our gradient map.

Although our disparity estimation technique is similar in spirit to [30], our use of gradients instead of image intensities is a notable departure from their approach. Furthermore, they employ several heuristics (e.g., repeated texture, lack of texture, outlier motion) to compute confidence, whereas our confidence estimates are based directly on the quadratic fit.

3.2. In-focus and defocus image distributions

The difference in sharpness between the background and reflection layers provides yet another valuable cue for layer separation. This idea was explored in [22]. The defocused reflection layer has fewer large gradients than the in-focus background. Following [22], we model the blurred reflection layer's gradient distribution using a Gaussian function with a narrow spread as

$$P_R(l) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{l^2}{2\sigma^2}}, \tag{4}$$

where l represents the gradient value, and σ denotes the standard deviation of the Gaussian.

It is well known that the gradients of natural images have a heavy-tailed distribution, and that this distribution can be modeled using a hyper-Laplacian function [16, 17]. Therefore, we express the probability distribution of the gradients of the in-focus background layer as

$$P_B(l) = e^{-\alpha|l|^p}, \tag{5}$$

where α is a positive scalar, and we set $p = \frac{2}{3}$ [16].

The work by [22] also applies different distributions on the gradients of the two layers. However, they model even the background's gradient distribution using a Gaussian. They then force the distribution to have a tail by applying the *max* operator and preventing the gradients from getting close to zero. In comparison, our use of the hyper-Laplacian distribution more naturally encourages large gradients in the background. Furthermore, [22] relies purely on relative smoothness, and so their method fails in cases where there is not a clear difference in sharpness between the two layers (see example 1 of Fig. 4). Our proposed method assimilates additional information about the gradients using disparity as a cue, and yields stable performance even when the reflection layer is only slightly out-of-focus.

3.3. Cost function

Our cost function uses a probabilistic model to seek the most likely explanation of the superimposed image using

Algorithm 1 Reflection removal using a dual-pixel sensor

Input: : Input image $\mathbf{g}$, the left $\mathbf{g}_{\text{LV}}$ and right $\mathbf{g}_{\text{RV}}$ DP views, relative weight λ, maximum iterations Q.

Output: : Background $\mathbf{b}$, and blurred reflection $\mathbf{r}$.

Compute $\mathbf{C}$ using $\mathbf{g}_{\text{LV}}$ and $\mathbf{g}_{\text{RV}}$ (see Section 3.1)
$\mathbf{D}_* = \mathbf{CD}$
$q = 0$
$\mathbf{b} = (\mathbf{D}^T\mathbf{D} + \lambda {\mathbf{D}_*}^T\mathbf{D}_*)^{-1}(\lambda {\mathbf{D}_*}^T\mathbf{D}_*\mathbf{g})$
do
  $e_i = \Big(\max\big(|(\mathbf{Db})_i|, 0.001\big)\Big)^{(p-2)}$
  $\mathbf{E} = \text{diag}(e_i)$
  $\mathbf{b} = (\mathbf{D}^T\mathbf{ED} + \lambda {\mathbf{D}_*}^T\mathbf{D}_*)^{-1}(\lambda {\mathbf{D}_*}^T\mathbf{D}_*\mathbf{g})$
  $q++$
while $q < Q$
$\mathbf{r} = \mathbf{g} - \mathbf{b}$

the probabilities of the background and reflection layers. Specifically, we maximize the joint probability $P(\mathbf{b}, \mathbf{r})$. Assuming that the background and the reflection are independent [19], the joint probability can be expressed as the product of the probabilities of each of the two layers – that is, $P(\mathbf{b}, \mathbf{r}) = P(\mathbf{b})P(\mathbf{r})$. Following [36], we define our distribution over both background and reflection layers using the histogram of derivative filters as

$$P(\mathbf{z}) \approx \prod_{i,k} P\big((\mathbf{D}_k\mathbf{z})_i\big), \quad \mathbf{z} = \text{ either } \mathbf{b} \text{ or } \mathbf{r}, \tag{6}$$

where we assume that the horizontal and vertical derivative filters $\mathbf{D}_k \in \{\mathbf{D}_x, \mathbf{D}_y, \mathbf{D}_{xx}, \mathbf{D}_{xy}, \mathbf{D}_{yy}\}$ are independent over space and orientation.

Maximizing $P(\mathbf{b}, \mathbf{r})$ is equal to minimizing its negative *log*, and from equations (4)(5)(6), we obtain the following cost function:

$$\arg\min_{\mathbf{b},\mathbf{r}} \left\{ \sum_{i,k} \left(|(\mathbf{D}_k\mathbf{b})_i|^p + \lambda\big((\mathbf{D}_k\mathbf{r})_i\big)^2 \right) \right\}, \tag{7}$$

where we integrate the relative weight between the two terms and the multiplier $\frac{1}{2\alpha\sigma^2}$ into a single parameter λ, which controls the amount of defocus blur in the reflection layer. This can be rewritten as

$$\arg\min_{\mathbf{b},\mathbf{r}} \Big\{ ||\mathbf{Db}||_p^p + \lambda||\mathbf{Dr}||_2^2 \Big\}, \tag{8}$$

where the matrix $\mathbf{D}$ consists of the five $\mathbf{D}_k$s vertically stacked. Expressing in terms of a single layer $\mathbf{b}$, and incorporating the confidence values $\mathbf{C} = \text{diag}(c_i)$ from equation (3) to enforce agreement with the labeled gradients, we obtain

$$\arg\min_{\mathbf{b}} \Big\{ ||\mathbf{Db}||_p^p + \lambda||\mathbf{CD}(\mathbf{g} - \mathbf{b})||_2^2 \Big\}. \tag{9}$$

Equation (9) can be solved using iterative reweighed least squares, and the steps are outlined in Algorithm 1. In all our experiments, the optimization converges quickly within a few iterations. Note that our cost function is based purely on gradients. Therefore, we finally rescale the recovered background and reflection images based on the input image's intensity range.

We would like to add that we chose not to include any explicit image fidelity terms based on the image formation model in equation (1) inside our cost function. The defocus blurring operation encoded by the matrices $\mathbf{W}_{\text{LV}}$ and $\mathbf{W}_{\text{RV}}$ can be space-varying depending on the depth of the reflected object. A per-pixel-varying defocus kernel is hard to reliably estimate from the composite image. Moreover, the blur size is a function of aperture (see equation 3 of [30]). Observe that our cost function based on gradients is not aperture-specific, does not entail complex per-pixel depth estimation, and is straightforward to optimize.

4. Experiments

There are no publicly available datasets for reflection removal that provide dual pixel data. Therefore, to evaluate the performance of our proposed algorithm, we capture our own dataset using a dual pixel camera. Although DP technology exists on most modern cameras, the vast majority of these devices do not provide users access to DP data. This is primarily because DP autofocus occurs at the very early stages of the camera pipeline and the current raw readout hardware combines the information to mimic a single readout for each pixel. As far as we are aware, there is no direct Camera2 API [1] call to read off the DP measurements even from the Google Pixel 2 phone used in [30]. As a result, we use the Canon EOS 5D Mark IV DSLR camera, one of the few commercially available cameras that provides access to sensor's DP data, to capture our dataset.

4.1. Data capture

Our dataset is divided into two categories – controlled indoor scenes with ground truth and scenes captured "in the wild". Following the data capture methodology adopted by the recent single-image reflection removal benchmark dataset of [31], we use different postcards as background and reflection (see Fig. 4) for the controlled dataset. We select postcards with texture ranging from medium to high for both background and reflection, and combine them pairwise in a manner that our dataset has a wide diversity of complex overlapping textures. In particular, we select six postcards for background and five postcards for reflection, for a total of 30 different scenes.

The defocus blur size and the disparity are functions of the aperture. To evaluate our algorithm's robustness to degree of defocus blur and extent of disparity, we also vary the aperture value. Specifically, we select five different aperture

Figure 4. Examples from our controlled dataset.

sizes {F13, F10, F8, F5.6, F4}. In the supplementary material, we provide animations that switch between the two DP views to better reveal how the defocus blur and disparity change with aperture. For each of the 30 scenes, we capture images using these five different apertures, giving us a total of 150 images for the controlled dataset. In order to make the controlled scenes even more challenging, we place a light source close to the postcard in front of the glass to boost the interference from the reflection [31]. The ground truth background layer is captured with the portable glass pane removed.

While a controlled setup allows for a quantitative evaluation of our proposed method as well as competing algorithms, these scenes do not necessarily reflect the complexities encountered in images captured in an unconstrained manner. Therefore, we augment our dataset with images captured in the wild (see Fig. 5 for some examples). For the in-the-wild category, we found it difficult to capture the ground truth (due to motion in the scene, the glass pane being fixed, etc.), and so we analyse results only qualitatively.

4.2. Comparisons

We compare our results with six contemporary reflection removal algorithms – four single-image algorithms, LB14 [22], WS16 [33], ZN18 [40], and YG18 [39], and two motion-based multi-image algorithms, LB13 [21], GC14 [11]. The codes for all six methods have been made publicly available by the authors. For the single-image algorithms, we use the default parameters mentioned in their paper or provided in the original code, and feed the captured image as input. We chose the conventional methods of LB14 [22] and WS16 [33] for comparison because they operate under the same assumptions that we do; the background is sharp and the reflection is defocused. YG18 [39] and ZN18

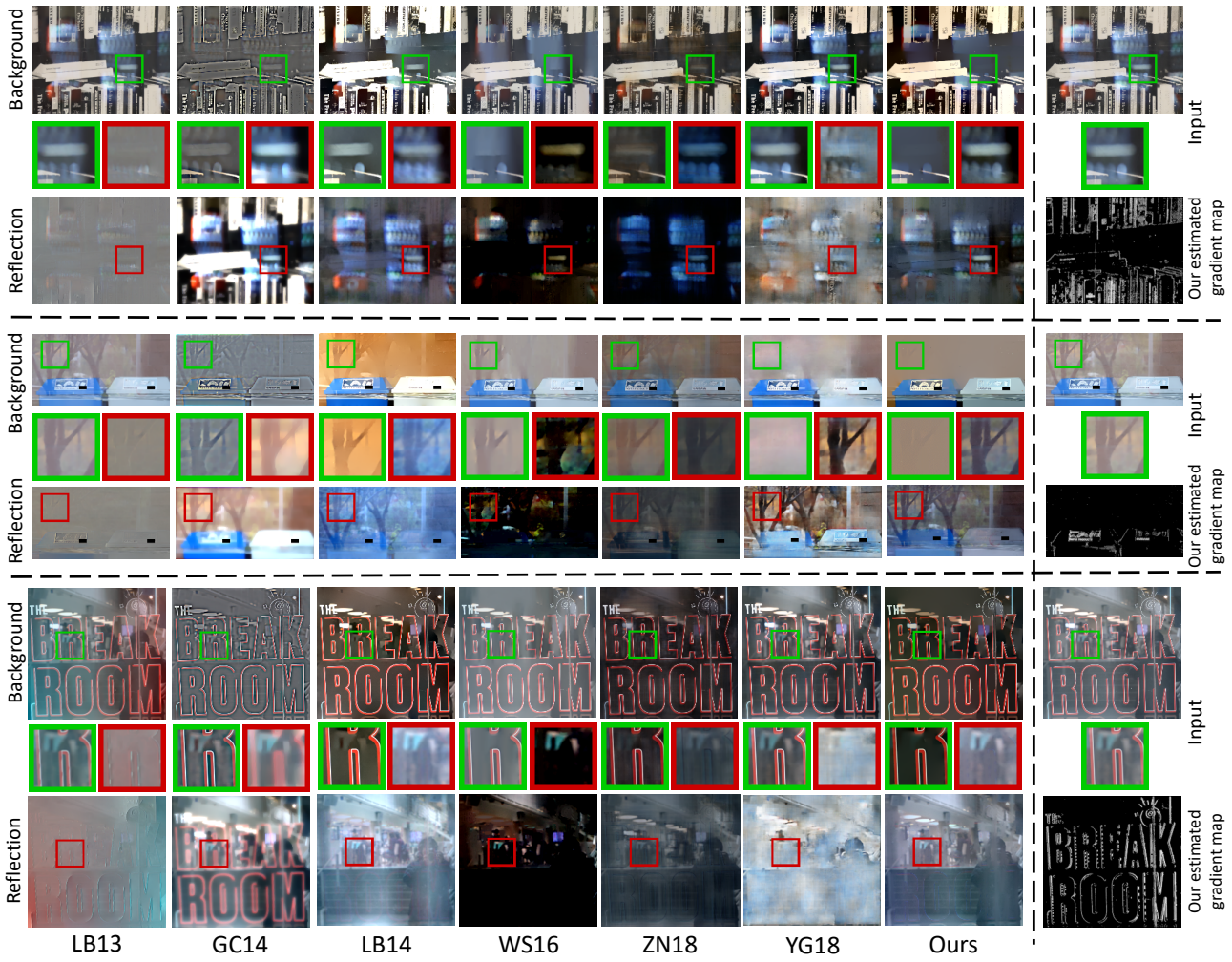


Figure 5. Examples from our in-the-wild dataset.

[40] are the two most recent deep learning methods for single-image reflection removal with state-of-the-art performance. Since the two sub-aperture views are available to us from the DP sensor, and these are essentially two different viewpoints of the scene, we also compare against the multi-image methods of LB13 [21] and GC14 [11] which exploit motion cues for layer separation. For a fair comparison against these methods, we restricted their search space to pure translation instead of a full homography. We provide the left and right DP views as input to the multi-image methods because the change in viewpoint is highest between these two images. In our experiments, including the input image along with the DP views did not improve their performance. Code for the light-field camera-based methods discussed in Section 1.1 is not publicly available.

4.3. Error metrics

We quantitatively compare the results of our proposed algorithm as well as competing techniques on the controlled dataset. We evaluate performance using several metrics: (i) peak signal to noise ratio (PSNR) and (ii) structural similarity index (SSIM) [35] are the two most commonly employed. Following [31], we also use (iii) local mean squared error as a similarity measure (sLMSE), (iv) normalized cross correlation (NCC), and (v) structure index (SI). Please refer to [31] for more details of metrics (iii) to (v).

4.4. Results on controlled scenes

The performance of LB13 [21], GC14 [11], LB14 [22], WS16 [33], ZN18 [40], YG18 [39], and our proposed method on the 150 images in the controlled category of our dataset for the five error metrics is recorded in Table 1. It can be observed that we outperform competing approaches by a sound margin on all metrics. Fig. 4 shows

Method	PSNR (dB)	SSIM	sLMSE	NCC	SI
LB13 [21]	16.12	0.689	0.870	0.966	0.758
GC14 [11]	16.02	0.798	0.888	0.945	0.496
LB14 [22]	14.20	0.842	0.797	0.981	0.840
WS16 [33]	16.62	0.836	0.884	0.975	0.837
ZN18 [40]	15.57	0.797	0.867	0.979	0.818
YG18 [39]	16.49	0.832	0.871	0.978	0.847
Ours	**19.45**	**0.883**	**0.946**	**0.982**	**0.870**

Table 1. Quantitative results on our controlled dataset.

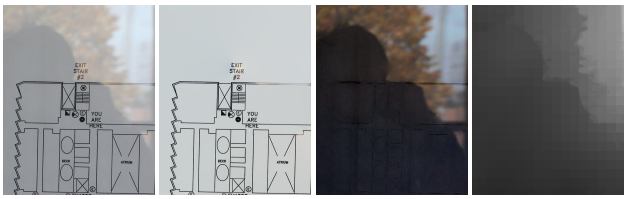
three representative examples from our controlled set with three different aperture settings. We noticed that the multi-image methods LB13 [21] and GC14 [11] do not perform well in general because both methods rely on large changes in viewpoint, whereas the baseline between the DP views is very narrow. The first row of Fig. 4 shows an example captured at F13 aperture value. Although the background does not have a lot of texture, the reflection is sharp due to the narrow aperture, and ZN18 [40] and YG18 [39] have traces of reflection in the top right of the image. It can be observed from the zoomed-in regions in the second row that LB14 [22] and WS16 [33] both also have residual reflection. In comparison, our method recovers both background and reflection (shown in the third row) more accurately.

Another example with a highly textured background as well as reflection captured at the F8 aperture is shown next. Competing techniques erroneously remove either too little (red box LB14 [22]) or too much (green box YG18 [39]) detail from the background, or miscalculate the overall contribution of the reflection layer. Our output more closely matches the ground truth when compared to other algorithms. The third example shot at the F4 aperture is more challenging because although the reflection is blurred, it covers a significant portion of the heavily textured background. All methods suffer from a loss of detail in this case. However, our method still produces a fairly good separation of the background and reflection layers.

4.5. Results on in-the-wild scenes

Fig. 5 shows three examples from our in-the-wild dataset. Since there is no ground truth, we provide zoomed-in regions corresponding to background (green) and reflection (red) for a visual comparison of various algorithms. Our estimated background gradient map is also shown. It can be observed that our method performs consistently well as opposed to competing techniques. More results are provided in the supplementary material.

We fix the parameters $\lambda = 100$ and $Q = 3$ for all experiments in this paper. On a 3.10 GHz processor with 32 GB RAM, our MATLAB algorithm takes approximately 2 minutes to process an 800×800 image.



(a) (b) (c) (d)

Figure 6. (a) Input image, (b) our estimated background, (c) our estimated reflection, and (d) depth map of the reflection layer.

An interesting extension of our work is the ability to recover a coarse depth map of the reflected scene. An example is demonstrated in Fig. 6. By subtracting out the estimated background from the left and right views, we can obtain the reflected scene as observed from the left and right views (see equation 1). These two images can then be used to extract a depth map of the reflected scene following the disparity estimation technique of Wadhwa et al. [30].

5. Discussion and summary

We have proposed a method to perform reflection removal by exploiting the data available on a DP sensor. We used the defocus-disparity cues present in the two sub-aperture views to simplify the task of determining which image gradients belong to the background layer. This well-labeled gradient map allows our optimization scheme to recover the background layer more accurately than other methods that do not use this additional information. The best part of our approach is that it does not require hardware modifications or training – instead it uses data already available within each camera shot. The only downside is most camera APIs currently do not provide access to this useful data. We hope this work will inspire manufacturers to provide access. In the meantime, we offer a new dataset for reflection removal that provides the two DP sub-aperture views.

We do note that our defocus-disparity cues are based on the assumption that the reflection layer is out of focus. Thus, one limitation of our approach is that we cannot fully distinguish between the gradients of the two layers if the background and the scene being reflected are at nearly equal distances from the glass – that is, both layers are in sharp focus, and the disparity is too small to be detected. One idea for future work is to use focus bracketing to combine multiple DP images for improved layer recovery.

Acknowledgment

This study was funded in part by the Canada First Research Excellence Fund for the Vision: Science to Applications (VISTA) programme and an NSERC Discovery Grant.

References

[1] Google, Inc.: Camera2 API Package Summary. http://developer.android.com/reference/android/hardware/camera2/. Accessed: 2016-07-16.

[2] A. Agrawal, R. Raskar, and R. Chellappa. Edge suppression by gradient field transformation using cross-projection tensors. In *CVPR*, 2006.

[3] Amit Agrawal, Ramesh Raskar, Shree K. Nayar, and Yuanzhen Li. Removing photography artifacts using gradient projection and flash-exposure sampling. *ACM Transactions on Graphics*, 24:828–835, 2005.

[4] Robert Anderson, David Gallup, Jonathan T Barron, Janne Kontkanen, Noah Snavely, Carlos Hernández, Sameer Agarwal, and Steven M Seitz. Jump: Virtual reality video. *SIGGRAPH Asia*, 2016.

[5] E. Be'ery and A. Yeredor. Blind separation of superimposed shifted images using parameterized joint diagonalization. *IEEE Transactions on Image Processing*, 17(3):340–353, 2008.

[6] Paramanand Chandramouli, Mehdi Noroozi, and Paolo Favaro. ConvNet-based depth estimation, reflection separation and deblurring of plenoptic images. In *ACCV*, 2016.

[7] Y. Chung, S. Chang, J. Wang, and S. Chen. Interference reflection separation from a single image. In *WACV*, 2009.

[8] Y. Diamant and Y. Y. Schechner. Overcoming visual reverberations. In *CVPR*, 2008.

[9] Qingnan Fan, Jiaolong Yang, Gang Hua, Baoquan Chen, and David Wipf. A generic deep architecture for single image reflection removal and image smoothing. In *ICCV*, 2017.

[10] K. Gai, Z. Shi, and C. Zhang. Blind separation of superimposed moving images using image statistics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(1):19–32, 2012.

[11] X. Guo, X. Cao, and Y. Ma. Robust separation of reflection from multiple images. In *CVPR*, 2014.

[12] B. Han and J. Sim. Glass reflection removal using co-saliency-based image alignment and low-rank matrix completion in gradient domain. *IEEE Transactions on Image Processing*, 27(10):4873–4888, 2018.

[13] Jinbeum Jang, Yoonjong Yoo, Jongheon Kim, and Joonki Paik. Sensor-based auto-focusing system using multi-scale feature extraction and phase correlation matching. *Sensors*, 15(3):5747–5762, 2015.

[14] Ole Johannsen, Antonin Sulc, and Bastian Goldluecke. Variational separation of light field layers. In *Vision, Modeling and Visualization*. The Eurographics Association, 2015.

[15] N. Kong, Y. Tai, and J. S. Shin. A physically-based approach to reflection separation: From physical modeling to constrained optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(2):209–221, 2014.

[16] Dilip Krishnan and Rob Fergus. Fast image deconvolution using hyper-laplacian priors. In *NIPS*, 2009.

[17] Anat Levin, Rob Fergus, Frédo Durand, and William T. Freeman. Image and depth from a conventional camera with a coded aperture. *ACM Transactions on Graphics*, 26(3), 2007.

[18] A. Levin and Y. Weiss. User assisted separation of reflections from a single image using a sparsity prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(9):1647–1654, 2007.

[19] Anat Levin, Assaf Zomet, and Yair Weiss. Learning to perceive transparency from the statistics of natural scenes. In *NIPS*, 2002.

[20] A. Levin, A. Zomet, and Y. Weiss. Separating reflections from a single image using local features. In *CVPR*, 2004.

[21] Y. Li and M. S. Brown. Exploiting reflection change for automatic reflection removal. In *ICCV*, 2013.

[22] Y. Li and M. S. Brown. Single image layer separation using relative smoothness. In *CVPR*, 2014.

[23] Y. Ni, J. Chen, and L. Chau. Reflection removal on single light field capture using focus manipulation. *IEEE Transactions on Computational Imaging*, 4:562–572, 2018.

[24] Bernard Sarel and Michal Irani. Separating transparent layers through layer information exchange. In *ECCV*, 2004.

[25] Yoav Y. Schechner, Nahum Kiryati, and Ronen Basri. Separation of transparent layers using focus. *International Journal of Computer Vision*, 39(1):25–39, 2000.

[26] Y. Y. Schechner, J. Shamir, and N. Kiryati. Polarization-based decorrelation of transparent layers: The inclination angle of an invisible surface. In *ICCV*, 1999.

[27] YiChang Shih, D. Krishnan, F. Durand, and W. T. Freeman. Reflection removal using ghosting cues. In *CVPR*, 2015.

[28] Przemysław Śliwiński and Paweł Wachel. A simple model for on-sensor phase-detection autofocusing algorithm. *Journal of Computer and Communications*, 1(06):11, 2013.

[29] Chao Sun, Shuaicheng Liu, Taotao Yang, Bing Zeng, Zhengning Wang, and Guanghui Liu. Automatic reflection removal using gradient intensity and motion cues. In *ACM Multimedia*, 2016.

[30] Neal Wadhwa, Rahul Garg, David E. Jacobs, Bryan E. Feldman, Nori Kanazawa, Robert Carroll, Yair Movshovitz-Attias, Jonathan T. Barron, Yael Pritch, and Marc Levoy. Synthetic depth-of-field with a single-camera mobile phone. *ACM Transactions on Graphics*, 37(4):64:1–64:13, 2018.

[31] R. Wan, B. Shi, L. Duan, A. Tan, and A. C. Kot. Benchmarking single-image reflection removal algorithms. In *ICCV*, 2017.

[32] Renjie Wan, Boxin Shi, Ling-Yu Duan, Ah-Hwee Tan, and Alex C. Kot. CRRN: multi-scale guided concurrent reflection removal network. In *CVPR*, 2018.

[33] R. Wan, B. Shi, T. A. Hwee, and A. C. Kot. Depth of field guided reflection removal. In *ICIP*, 2016.

[34] Qiaosong Wang, Haiting Lin, Yi Ma, Sing Bing Kang, and Jingyi Yu. Automatic layer separation using light field imaging. *arXiv*, 2015.

[35] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on image processing*, 13(4):600–612, 2004.

[36] Y. Weiss. Deriving intrinsic images from image sequences. In *ICCV*, 2001.

[37] Tianfan Xue, Michael Rubinstein, Ce Liu, and William T. Freeman. A computational approach for obstruction-free

photography. *ACM Transactions on Graphics*, 34(4):79:1–79:11, 2015.

[38] Q. Yan, Y. Xu, X. Yang, and T. Nguyen. Separation of weak reflection from a single superimposed image. *IEEE Signal Processing Letters*, 21(10):1173–1176, 2014.

[39] Jie Yang, Dong Gong, Lingqiao Liu, and Qinfeng Shi. Seeing deeply and bidirectionally: A deep learning approach for single image reflection removal. In *ECCV*, 2018.

[40] Xuaner Zhang, Ren Ng, and Qifeng Chen. Single image reflection separation with perceptual losses. In *CVPR*, 2018.