The following publication T. Li, D. P. K. Lun, Y. Chan and Budianto, "Robust Reflection Removal Based on Light Field Imaging," in IEEE Transactions on Image Processing, vol. 28, no. 4, pp. 1798-1812, April 2019 is available at https://doi.org/10.1109/TIP.2018.2880510.

Robust Reflection Removal Based on Light Field Imaging

Tingtian Li, Daniel P.K. Lun, Yuk-Hee Chan, and Budianto

***Abstract*—In daily photography, it is common to capture images in the reflection of an unwanted scene. This circumstance arises frequently when imaging through a semi-reflecting material such as glass. The unwanted reflection will affect the visibility of the background image and introduce ambiguity that perturbs the subsequent analysis on the image. It is a very challenging task to remove the reflection of an image since the problem is severely ill-posed. In this paper, we propose a novel algorithm to solve the reflection removal problem based on light field (LF) imaging. For the proposed algorithm, we first show that the strong gradient points of an LF epipolar plane image (EPI) are preserved after adding to the EPI of another LF image. We can then make use of these strong gradient points to give a rough estimation of the background and reflection. Rather than assuming that the background and reflection have absolutely different disparity ranges, we propose a sandwich layer model to allow them to have common disparities, which is more realistic in practical situations. Then, the background image is refined by recovering the components in the shared disparity range using an iterative enhancement process. Our experimental results show that the proposed algorithm achieves superior performance over traditional approaches both qualitatively and quantitatively. These results verify the robustness of the proposed algorithm when working with images captured from real-life scenes.**

***Index Terms*—Light field, reflection removal, image separation**

I. Introduction

Images with reflections of unwanted scenes are frequently obtained in daily photography activities, such as taking pictures of a painting protected by glass, imaging the outside scenery through the window of a room, video recording the movement of underwater objects from above the water surface, and so on. It is important to remove the unwanted reflection since it not only affects the visibility of the background but also introduces ambiguity that perturbs the subsequent analysis of the image. Traditionally, photographers make use of a polarizing filter to reduce the reflection. However, a polarizing filter can remove the reflected component completely only when the viewing angle is equal to the Brewster angle [1]. If the reflection comes from various directions, the reflected image will still be visible after the polarizing filter is used.

Alternatively, a reflection can also be removed using image processing approaches. An image I with the background image I_B and reflection I_R can be mathematically modeled as follows:

$$I = I_B + I_R. \tag{1}$$

Directly solving I_B or I_R from I is a typical blind image separation problem, which is known to be severely ill-posed [2]. Since two variables must be solved from only one equation, various priors are required for solving this unconstrained problem. Most priors that the previous methods adopted are gradient based, such as gradient sparsity and gradient uncorrelation [3-6]. The former is a well-known natural image property, and the latter is based on the observation that the strong gradients of I_B and I_R are normally non-overlapped. However, the effect of adding only these priors in the optimization process is limited due to the enormous variation of natural images. Researchers tend to utilize multiple images to remove reflections. Given a sequence of images with reflections, the relative motion between I_B and I_R among those images can be exploited for their separation. Different approaches based on two-dimensional (2D) homography [4], scale-invariant feature transform (SIFT) flow [6, 7], and optical flow [5] are proposed. The multiple-image-based methods have demonstrated better performance than the single-image methods. However, these multiple-image-based methods all have strong assumptions on the ways that they utilize motion, as will be discussed in Section II. In addition, all of them require multiple shots of the target scene and hence are not suitable for dynamic applications in which either the background or reflection objects are moving.

In contrast to traditional cameras, light field (LF) cameras can capture multiple views of a scene in one exposure. Due to the commercialization efforts of Lytro and Raytrix, currently, people can easily acquire an LF camera at a reasonable cost. Four-dimensional (4D) LF imaging [8] has demonstrated its power in solving various problems, such as refocusing [9-11], depth estimation [12-15] and super-resolution imaging [13, 16], in the computer vision area. Quite recently, LF cameras have been also used to solve the reflection removal problem [17]. By assuming that the background and reflection are at two absolutely different distances from the camera, the method in [17] applies a fixed threshold to separate the background and reflection pixels with respect to their disparities. Such an assumption, however, is not valid in many practical situations. First, some of the background and reflection components can

This work was fully supported by the Hong Kong Polytechnic University under research grant RU9P.

Tingtian Li, Daniel P.K. Lun, Y.H. Chan and Budianto are with the Department of Electronic and Information Engineering, The Hong Kong Polytechnic University (e-mail: tingtianpolyu.li@connect.polyu.hk; enpklun@polyu.edu.hk; enyhchan@polyu.edu.hk; budianto@ieee.org).

share the same disparity range. Second, for those background and reflection components that are close to each other in disparity, the errors in the disparity estimations can lead to errors in their classification. In this paper, we first explore the LF epipolar plane image (EPI) [13, 18] and show that its strong gradient points will be preserved after adding to the EPI of another LF image. This property lets us easily identify the strong gradient points of the background and reflection images, and we can further use them to obtain a rough estimation of each image layer by a sparse regularization process. To solve the problem that the background and reflection images can share the same disparity range, we propose a sandwich layer model that allows the background and reflection images to have components that share the same disparity range. We then propose a method that gradually refines the initial background estimate by detecting and recovering the components of the background in the shared disparity range. It is achieved based on an observation that the strong gradient points of these components can be found in both the initial background estimate and its residue. It gives us the clue to recover these components and add them back into the initial background estimate.

To summarize, the main contributions of this paper are as follows:

1) We explore the theoretical support of using LF EPI to estimate the disparities of the different layers of a superimposed LF image. We verify that if an LF image is formed by the superimposition of two LF image layers of different disparities, the EPI strong gradient points of both images will be at different positions of the combined EPI, and the gradient values will be preserved.

2) We propose a general sandwich model to describe the depth range of the background and reflection images. This model allows a shared depth range for both images, which is more realistic in practical situations. Following this model, the proposed method does not require the background and reflection images to have absolutely different depth ranges as in the existing approach.

3) We develop a new method to detect and extract the components of the background images that have the same disparities as the reflection. It is achieved based on an observation that these components can be found in both the initial background estimate and its residue.

This paper is organized as follows: After the introduction in Section I, we list some related studies on reflection removal in Section II. In Section III, we analyze and illustrate how the LF EPI gradients can be used to classify different layers of a superimposed LF image. In Section IV, we present the proposed method for detecting and recovering the background components in the shared disparity region and use them to iteratively enhance the initial background estimate. The experimental and simulation results are shown in Section V. Last, we draw the conclusions in Section VI.

II. Related Work

Taking images through semi-reflective material is a tricky task for many photographers because an unwanted reflection of another scene will be added to the captured image. As mentioned in Section I, the traditional method of using a polarizing filter is insufficient because reflection can come from different angles [19]. In [20], it is suggested that image reflection can be removed by combining images taken with and without flash. This approach, however, works only for weak reflections in which the flash light is strong enough to overwhelm them. In addition, the flashlight can also introduce highlights and other artifacts to the final image. Rather than use extra optical devices, researchers have also looked for pure image processing solutions. However, since the reflection image often has similar morphological properties as the desired image, it cannot be easily removed by using traditional morphological filters [21]. In [3, 22, 23], the statistical properties of natural images are investigated with regard to separating the background and reflection from a single image. Reference [3] assumes that the gradients of each image layer follow a mixture of two Laplacian distributions, and it requires users to manually label the strong gradients of each layer to reduce the ambiguities. However, the gradient distribution assumption is too simple for general natural images, and the requirement of user assistance lowers its practical value. In [22], it is assumed that the captured image's background layer is in focus and the reflection layer is defocused. Thus, it uses long-tailed and short-tailed distributions to model the gradients of the background and reflection layers, respectively. This gradient model is still not sufficient to describe all of the natural images, and the assumption that the background is in focus and the reflection is defocused is not always followed.

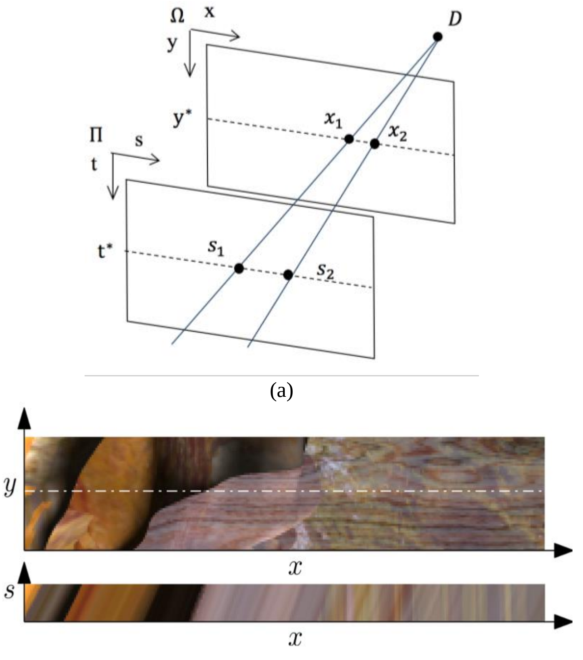
It appears to be very difficult to solve this severely ill-posed problem with only a single image. Therefore, recently, many multiple-image-based methods have been developed [4-6, 24]. These approaches usually assume that the background and reflection have quite distinct depth ranges in such a way that they can be separated based on the disparities evaluated from the multiple images taken from different angles of the scene. In [4], the 2D homography method is used to register the background. Since the background and reflection are assumed to have different disparities, using the homography of the background to align the images will blur the reflection. Then, a low rank decomposition method is applied to separate the background and the reflection. However, the method will only work if the background is a plane. If the background also has a disparity range, some parts of the background will also be blurred in the process. In [6], the SIFT flow is used to register the dominant background gradient. Since the reflection will fail to register due to its weak intensity, the gradients can be separated according to the extents of the alignments. The background can then be reconstructed from the gradients. Due to the weak reflection assumption, the method will fail if the reflection has a strong intensity, which is difficult to avoid in real scenes. In [5], different optical flows of the background and reflection images are adopted for their separation. This method, however, can easily fall into a local minimum because too many variables are regularized simultaneously. A very accurate initial guess is needed to guide the optimization process to the desired solution. In fact, for all of these approaches, multiple images are obtained sequentially, and hence, they are not suitable to dynamic scenes in which either the background or reflection objects are moving. Since LF cameras can obtain multiple-view images at one shot, they have been adopted for the reflection removal problem recently [17]. In that method, it is assumed that the background and reflection layers are at opposite sides of the actual sensor normal, in such a way that their slopes in the EPI will have opposite signs. Then, a fixed disparity threshold (zero in this case) can be set to identify the background and reflection gradients. This assumption, however, introduces a stringent requirement to the position of the background and reflection layers. One of them, but not both, must be within 1.5 m from the camera (assuming Lytro ILLUM is used). In addition, the camera must be pointed perpendicular to the reflecting surface, which is difficult because it relates to the style of the photography. In fact, in many situations, the reflection and the background images cannot be totally separated in terms of their distance to the camera. They can share the same disparity range. In addition, the errors in the disparity estimation further complicate the separation process. A more robust approach is required to solve the reflection removal problem without the abovementioned restrictions.

III. Using the EPI Gradient In Separating the Background and Reflection

In this section, we first provide a brief review on LF EPI and explain how its gradient can be used in the estimation of the disparity maps for the problem of reflection removal. Then, a new sandwich model is introduced for background and reflection layer classification.

A. Layer classification based on the EPI strong gradient points

Although there are several approaches to representing the light field, the 4D approach, which uses two planes to represent the viewpoints and image plane, as shown in Fig. 1(a), is the most popular [8]. In this figure, planes Π and Ω are the viewpoint plane and image plane, respectively. Here, we use the coordinate systems (s,t) for Π and (x,y) for Ω. Therefore, we can describe each light ray by a 4D coordinate (s,t,x,y). When we fix t and y as t^* and y^*, and we let s and x move, we will obtain the so-called EPI slice Σ_{y^*,t^*}. An example of LF EPI is shown in Fig. 1(b) (lower one). It can be seen that an EPI is formed by stripes of different orientations. This arrangement occurs because the distances between the adjacent views of an LF image are the same. For a pixel in the reference view, the corresponding pixels in the other views will be located at positions that are linearly proportional to the disparity. Thus, when showing them in the EPI, these pixels become a strip of a particular orientation as determined by the disparity. More specifically, the slope reciprocal $\Delta x/\Delta s$ that measures the pixel shift at $\Sigma_{y^*,t^*}(x,s)$ can represent the disparity at point (x,y^*) for the view (s,t^*) [13, 26]. Hence, the EPI slope is often used to evaluate the disparity and, in turn, the depth of an LF image. The slope directions can be obtained using the structure tensor [13, 27, 28], which determines the gradient direction by finding the eigenvectors in which the direction of the magnitude changes most rapidly or most slowly. The structure tensor for EPI Σ_{y^*,t^*} can be described as



(a)

(b)

Fig. 1. (a) The 4D LF model. A light ray can be described using the 4D coordinate (s,t,x,y). (b) The EPI of a slice of an LF image [25].

$$\begin{aligned} &J_{\Sigma_{y^*,t^*}}(x,s) \\ &= \begin{bmatrix} G_\sigma ** \big(\partial(x)\partial(x)\big) & G_\sigma ** \big(\partial(x)\partial(s)\big) \\ G_\sigma ** \big(\partial(x)\partial(s)\big) & G_\sigma ** \big(\partial(s)\partial(s)\big) \end{bmatrix} \\ &= \begin{bmatrix} J_{xx} & J_{xs} \\ J_{xs} & J_{ss} \end{bmatrix}, \end{aligned} \tag{2}$$

where $\partial(x)$ and $\partial(s)$ represent the gradient components in the x and s directions, respectively, at point (x,s) in EPI Σ_{y^*,t^*}. G_σ is a Gaussian kernel with variance σ, and the symbol $**$ denotes convolution. The disparity values at (x,y) for a reference view (s^*,t^*) can be obtained from the EPI Σ_{y,t^*} by [28]:

$$P_{\Sigma_{y,t^*}}(x,s^*) = \frac{\Delta x}{\Delta s^*} = \tan\theta, \tag{3}$$

where

$$\theta = \frac{1}{2}\arctan\left(\frac{J_{s^*s^*}-J_{xx}}{2J_{xs^*}}\right). \tag{4}$$

A reliability measure can also be generated as follows:

$$r_{\Sigma_{\mathrm{y},t^*}}(x,s^*) = \frac{\left(J_{s^*s^*}-J_{xx}\right)^2+4\left(J_{xs^*}\right)^2}{\left(J_{s^*s^*}+J_{xx}\right)^2}. \tag{5}$$

A disparity map $P_{\Sigma_{y,t^*}}(x,s^*)$ and reliability map $r_{\Sigma_{\mathrm{y},t^*}}(x,s^*)$ based on the EPIs in the horizontal direction can then be obtained by repeating the above for all y. We can also obtain a disparity map $P_{\Sigma_{x,s^*}}(y,t^*)$ and reliability map $r_{\Sigma_{x,s^*}}(y,t^*)$ based on the EPIs in the vertical direction using a similar approach. Then, the final disparity map for the reference view (s^*,t^*) is generated by selecting the disparity value that has higher reliability. In other words, we have

$$\begin{aligned} &P(x,y) = \\ &\begin{cases} P_{\Sigma_{y,t^*}}(x,s^*) & if\ r_{\Sigma_{y,t^*}}(x,s^*) > r_{\Sigma_{x,s^*}}(y,t^*) \\ P_{\Sigma_{x,s^*}}(y,t^*) & otherwise \end{cases}. \end{aligned} \tag{6}$$

In practice, if the reliability value is too small, then $P(x, y)$ can be inaccurate and will simply be set as *invalid*. One of the situations in which this arrangement will occur is when the pixel (x, y) has no gradient or a very weak gradient. Hence, $P(x, y)$ can also be considered to be the disparity map at the strong gradient points of the image.

For the problem of reflection removal, a reflection image is superimposed on the background image. When the scene is captured by an LF camera, the resulting EPIs will also be a superimposition of the EPIs of both images. Since these images can have different depths, we can find that the resulting EPI also has slopes of different angles, and they cross each other randomly in the EPI. Especially in the regions where they cross each other, it is difficult to determine the slope of the EPI pixels and classify them into the background or reflection layer. To address this problem, we consider again the gradient of the EPIs, of which the disparity map is derived in (2) to (6). In particular, we investigate the behavior of the strong and weak gradient points of the background and reflection as follows:

Case 1: Strong gradient points of both layers

This case is illustrated in Fig. 2(a) to (c). In the figure, both EPIs have two strong gradient points. When the EPIs are added up, the strong gradient points do not overlap each other, and they preserve the same values, as shown in Fig. 2(c). Such a phenomenon is not coincidental. It is known that the strong gradient points of an EPI correspond to the strong gradient points of the image. Due to the gradient independence assumption [3, 4], it is rare to have strong gradient points of two uncorrelated images overlapping each other. Consequently, we can also assume that the EPI strong gradient points of two uncorrelated images will be at different positions in the combined EPI. In addition, as shown in Fig. 2(c), the gradient value will remain the same wherever a strong gradient point is located in the combined EPI. Consequently, we can easily estimate the disparities at these strong EPI gradient points. An example is shown in Fig. 3. In this example, two LF images are added together with the weightings of 0.6 and 0.4, respectively. The central view of the resulting LF image is shown in Fig. 3(c). The EPIs of the resulting LF image are then generated. Based on the EPIs, we first estimate the disparity map of the image in Fig. 3(c) using the structure tensor method in (2) to (6), and we keep only those at the strong gradient points. It can be seen in Fig. 3(d) that the disparities of the two layers at the strong gradient points can be easily identified because they are at different positions.

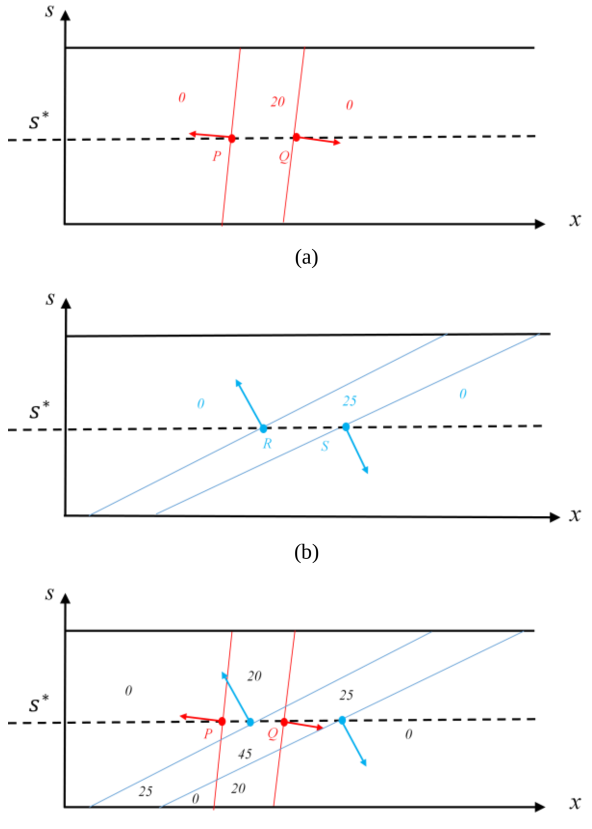


Fig. 2. An illustration of the relationship of the strong gradient points in the original and combined EPIs. (a) An EPI of an LF image. The numbers represent the pixel magnitudes in those regions. Two strong gradient points P and Q are noted. (b) An EPI of another LF image. Two strong gradient points R and S are noted. (c) The combined EPI. The numbers represent the pixel magnitudes after combination. It can be seen that all strong gradient points in (a) and (b) are located at different positions with the same values as before.

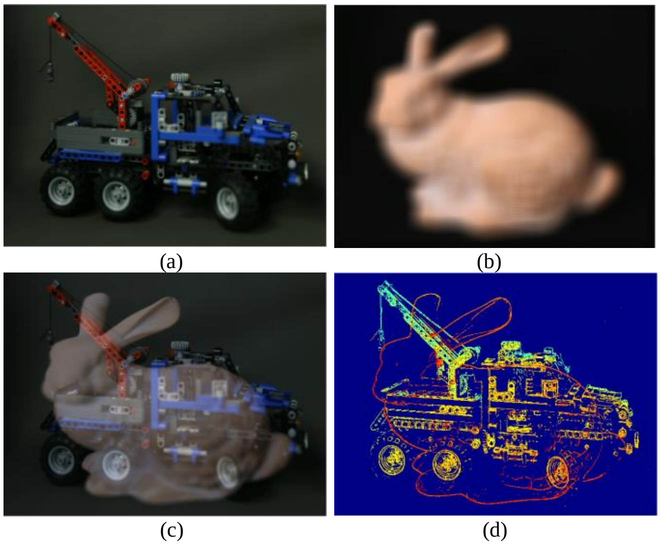


Fig. 3. (a) An LF image with all views overlapped. (b) Another LF image with all views overlapped. The extent of blurring represents the amount of disparity. We can see that the disparity of (b) is larger than that of (a). (c) The central view of an LF image generated by combining (a) and (b) with the weightings of 0.6 and 0.4, respectively. (d) The estimated disparity map based on the strong gradient points of the EPI of (c). The red and blue color means large and small disparities, respectively. Since in most cases they are not overlapped, they can be easily identified.

Case 2: Relatively weak gradient points of both layers

For the relatively weak EPI gradient points of both layers, they might or might not overlap with the EPI gradient points of the other layer. For those that do not overlap with another EPI gradient point, the disparity at those points can still be estimated as usual. In case they overlap with another EPI gradient point, their correct gradient value can no longer be recovered. They will be similar to those extremely small gradients that have no

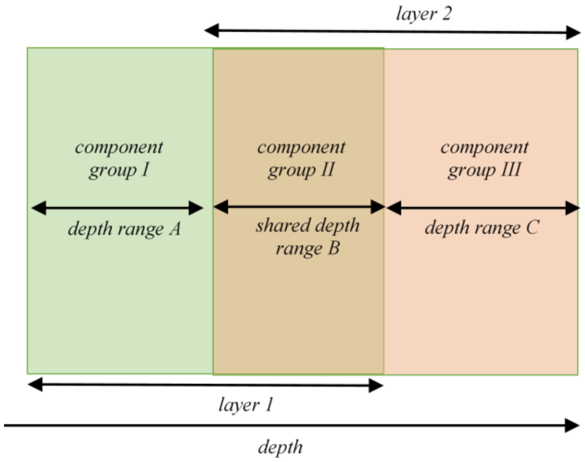


Fig. 4. The new sandwich model. In this model, component group I belongs to only layer I, and component group III belongs to only layer II (layer I is closer to the camera). Both layers share component group II.

definite directions. The disparity value estimated based on these gradients will appear as noise in the disparity map and will be regulated in the later optimization process.

To summarize, as the first step of our proposed algorithm, we make use of the structure tensor method in (2) to (6) to generate a disparity map in the EPI domain. Since the gradient points in the EPI domain have a close relationship with the gradient points in the spatial domain, the resulting disparity map will contain accurate disparity values at the strong gradient points of both the background and reflection images, and at the weak gradient points in case they do not overlap with other gradient points. We also expect that there will be noise caused by the overlapped weak gradient points of both images.

B. The sandwich model and initial image reconstruction

If the background and reflection have absolutely different depth ranges, the disparity map generated in Part A should be sufficient to classify most of the gradient points; additionally, we can use these gradient points to reconstruct the background and reflection images. Unfortunately, it is not uncommon in many practical situations that some components of the background and reflection to share a common depth range. This circumstance means that their disparities can be very similar. In addition, the disparity maps generated from light field images can also have errors. Purely using the disparity map to separate the background and reflection layers can be erroneous, especially for those components that have similar depth values. For this reason, we propose a new sandwich model, as shown in Fig. 4, to take care of such situations. As shown in the figure, the model has one shared depth range for both layers. Assume that we can find two thresholds, K_1 and K_2, which are at the boundaries of component groups I and II, as well as groups II and III, respectively. Then, all gradient points with disparities smaller than K_1 will belong to layer 1, and those greater than K_2 will belong to layer 2. For those that are greater than K_1 but smaller than K_2, it is difficult to classify them by only their disparities due to the reasons mentioned above. We will discuss in the next section how these components can be classified by exploring their relationships with the components in groups I and III.

To find the thresholds K_1 and K_2, we apply the K-means clustering method [29] (where K=2 in this case) on the estimated disparity values at all strong gradient points. We denote the centers of the two clusters as C_1 and C_2 ($C_1 < C_2$). Then, we set the two thresholds as

$$\begin{aligned} K_1 &= C_1 + \sigma \cdot (C_2 - C_1); \\ K_2 &= C_2 - \sigma \cdot (C_2 - C_1), \end{aligned} \tag{7}$$

where σ is a parameter to control the purity of the classification result. In our experiment, we set $\sigma = 0.2$, which is a conservative choice to ensure that the classification has a high true positive rate. Then, we only need to take care of those misclassified gradient points. Based on K_1 and K_2, two initial gradient masks are obtained as follows:

$$\begin{aligned} E_B^0 &= \{P(x,y) > K_2\,, \forall x, \forall y\}; \\ E_R^0 &= \{P(x,y) < K_1\,, \forall x, \forall y\}, \end{aligned} \tag{8}$$

where $P(x,y)$ is defined in (6). $E_B^0, E_R^0 \in \{0,1\}$ are the two initial gradient masks for the background and the reflection layers, respectively. Without loss of generality, we assume that the background layer is the closer layer; otherwise, only a change of symbols is required. Fig. 5(i) shows an example of the initial gradient masks. It can be seen in Fig. 5(i)(b) that the locations of some of the strong gradient points of the background image are correctly indicated in E_B^0. However, we can also find that some strong gradient points of the background miss out in E_B^0. Based on the masks, we can reconstruct the background and reflection images in the gradient domain as follows:

$$\begin{aligned} I_B^0 = \arg\min_{I_B^0} J = \|D * I_B^0\|_1 + \left\|D * I_{\bar{B}}^0\right\|_1 + \\ \lambda\left\|E_B^0 \cdot D * I_{\bar{B}}^0\right\|_1 + \lambda\left\|E_{\bar{B}}^0 \cdot D * I_B^0\right\|_1; \\ \text{s.t. } I_{\bar{B}}^0 = I - I_B^0;\ \ E_{\bar{B}}^0 = \mathbf{1} - E_B^0, \end{aligned} \tag{9}$$

$$\begin{aligned} I_R^0 = \arg\min_{I_R^0} J = \|D * I_R^0\|_1 + \left\|D * I_{\bar{R}}^0\right\|_1 + \\ \lambda\|E_R^0 \cdot D * I_R^0\|_1 + \lambda\left\|E_{\bar{R}}^0 \cdot D * I_R^0\right\|_1; \\ \text{s.t. } I_{\bar{R}}^0 = I - I_R^0;\ \ E_{\bar{R}}^0 = \mathbf{1} - E_R^0, \end{aligned} \tag{10}$$

where **1** refers to an all ‘1’ matrix, and λ is a constant. In (9) and (10), the initial estimates of the background and reflection image, i.e., I_B^0 and I_R^0, are obtained by minimizing the sum of a few sparsity priors in the gradient domain. This approach is based on the gradient sparsity assumption that the total gradient of the background (or reflection) should be sparser than that of the original image, which contains the sum of the background and reflection. Thus, when the estimate I_B^0 (or I_R^0) approaches to the true background (or reflection), its total gradient should approach the minimum. The same is applied to their residues $I_{\bar{B}}^0 = I - I_B^0$ and $I_{\bar{R}}^0 = I - I_R^0$. In addition, based on the gradient independence assumption [3, 4], the total gradient of the background after multiplying with the gradient mask of the reflection should be small since their strong gradient points will not overlap. Thus, when the estimate I_B^0 (or I_R^0) approaches the true background (or reflection), its total gradient after multiplying with the mask of its residue, which approaches the true reflection (or background), should approach the minimum. In (10), $D \equiv D_{i=1,\ldots,5}$ represents a set of derivative filter kernels such that $D_1 = D_2^T = [1, -1]$ are the first-order derivative filters in the horizontal and vertical directions, respectively; $D_3 = D_4^T = [1, -2, 1]$ and $D_5 = D_2 * D_1$ are the second-order derivative filters in the horizontal, vertical and diagonal directions, respectively. The use of the second-order filters is for rectifying the discontinuities in the gradient domain due to the rare situations in which the strong gradients overlap each other. Here, (9) and (10) can be solved by the iteratively reweighted least squares (IRSL) method. Fig. 5(ii) shows an example of the initial separation results. For ease of visualization, the biases of the resulting images are adjusted to the original biases as follows:

$$I_{display} = I_{result} - mean(I_{result}) + mean(I) \qquad (11)$$

In Fig. 5(ii), it can be seen that almost all of the components of the initial background estimate belong to the background layer. However, many components are missing and can be found in its residue. The same is applied to the initial reflection estimate. To enhance the separation results, we develop a new method to detect and recover the missing components from the residues, which will be described in the next section.

IV. Detect and Recover the Missing Components

As mentioned in Section III above, there can be components of both the background and reflection layers that share the same disparity range (i.e., component group II in Fig. 4). These components are supposed to be cut away from I_B^0 and I_R^0. It can be seen in the initial estimation result (Fig. 5(ii)(b) and (d)) that large parts of I_B^0 and I_R^0 are darkened. They are the parts that have been cut away. To retrieve back these missing components, we have another observation about the strong gradient points in the initial estimation. By comparing between I_B^0 and its residue $I_{\bar{B}}^0$ (such as Fig. 5(ii)(b) and (c)), we observe that although the background components in the shared depth range are supposed to be cut due to the conservative thresholds used in (8), the strong gradient points of the missing background components can still be visualized in I_B^0 (circled in Fig. 5(ii)(b)). This finding is due to the first two terms in (9), which promote the sparsity in the image. However, their magnitudes are rather small, such that directly detecting them based on their magnitude can be erroneous. Note that both I_B^0 and $I_{\bar{B}}^0$ contain the strong gradient points of the background's missing components, although the points in $I_{\bar{B}}^0$ are much clearer. On the other hand, the strong gradient points of the reflection image are less visualized in I_B^0. This finding occurs because the magnitude of the reflection is often much lower than the background because most semi-reflective materials such as glass can only partially reflect the light projected onto them. Thus, for a particular spatial position (x, y), if the gradients $I_B^0(x, y)$ and $I_{\bar{B}}^0(x, y)$ are the same, they likely belong to the background. Based on the same argument, if the gradients $I_R^0(x, y)$ and $I_{\bar{B}}^0(x, y)$ are the same, then they likely belong to the reflection. We will make use of this property to detect and recover the missing components in $I_{\bar{B}}^0$.

As mentioned above, directly detecting the gradients of the missing components in I_B^0 based on their weak magnitudes can be erroneous. Therefore, we suggest considering also the gradient directions. While there are several ways to detect the directions of the gradients, we suggest considering the histogram of oriented gradients (HOG) method [30]. HOG is a feature descriptor for object detection. It contains the weighted (according to the magnitude) distribution (histograms) of the directions of the gradients (oriented gradients) of an image cell that is normalized with the nearby cells within a block. It is suitable in this problem because HOG is invariant to the local illumination of the image and can measure the direction of the gradients that have small magnitude. The procedure is as follows. First, to avoid the disturbance from the very weak gradients whose orientations are very unstable, we only consider the gradients at some spatial position set $\varphi^t = \{(x, y) || \partial_B^t(x, y)| > \epsilon\}$, where $|\partial_B^t(x, y)|$ is the magnitude of the gradient of $I_B^t(x, y)$ at iteration t, and ϵ is a very small constant. Then, we compute the HOG feature vectors H_B^t, $H_{\bar{B}}^t$ and H_R^0 at each spatial position in the set $\varphi^t$of I_B^t, $I_{\bar{B}}^t = I - I_B^t$

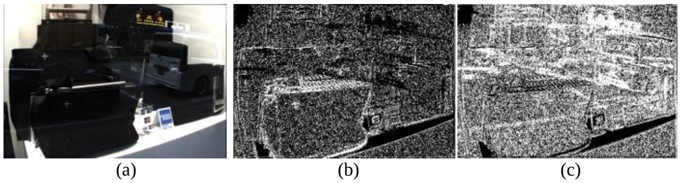


Fig 5(i). An example of E_B^0 and E_R^0. We can see that E_B^0 and E_R^0 can roughly separate the background and reflection gradient components.

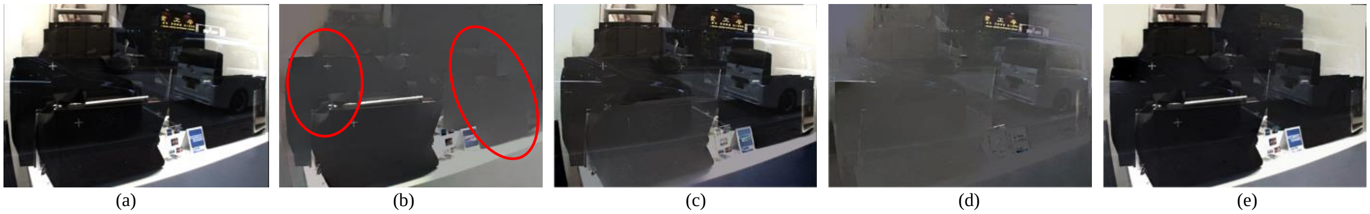


Fig. 5(ii). The initial separation results. All of the results are normalized by (11) for ease of visualization. (a) The original image I. (b) The initial estimate of the background layer I_B^0 . (c) The residue of the initial background estimate $I_{\bar{B}}^0 = I - I_B^0$. (d) The initial estimate of the reflection layer I_R^0. (e) The residue of the initial reflection estimate $I_{\bar{R}}^0 = I - I_R^0$. It can be seen that the components of I_B^0 almost only belong to the background layer, and its residue $I_{\bar{B}}^0$ contains not only the reflection components but also the missing background components. And similarly, I_R^0 loses some reflection components. which can be found in its residue $I_{\bar{R}}^0$.

and I_R^0, respectively. To retain the spatial resolution, we use a relatively small cell size of 3×3, and the block size is 2×2, as usual. Here, we use the UoCTTI variant [31, 32] of HOG, of which the feature vector length for every pixel is 31. Thus, the size of every feature vector is $h \times w \times 31$, where h and w are the height and width of the image. Then, we measure the Euclidean distances of H_B^t and $H_{\bar{B}}^t$ as well as H_R^0 and $H_{\bar{B}}^t$, as follows:

$$\begin{aligned} U_B^t(x,y) &= \left\|H_B^t(x,y) - H_{\bar{B}}^t(x,y)\right\|_2 ; \\ U_R^t(x,y) &= \left\|H_R^0(x,y) - H_{\bar{B}}^t(x,y)\right\|_2 , \end{aligned} \tag{12}$$

for all $(x,y) \in \varphi^t$. U_B^t measures the similarity between the HOG in I_B^t and its residue $I_{\bar{B}}^t$. If $U_B^t(x,y)$ is small, then the gradient at (x,y) of I_B^t and $I_{\bar{B}}^t$ should belong to the background, as discussed above. $U_R^t(x,y)$ measures the similarity between the gradients in I_R^0 and $I_{\bar{B}}^t$. Due to the conservative thresholds used in (7), I_R^0 contains mainly the components of the reflection layer. Additionally, $I_{\bar{B}}^t$ has the components of the reflection layer. Thus, if $U_R^t(x,y)$ is small, then the gradient at (x,y) of $I_{\bar{B}}^t$ should belong to the reflection. Then, U_B^t at the same point (x,y) should be large, since I_B^t should not have reflection components. Thus, U_R^t can be used to validate U_B^t in the classification process.

To perform the classification, we borrow the ideas of the Markov Random Field (MRF) [33] and the K-nearest neighbors (KNN) matting [34] to formulate the following optimization function:

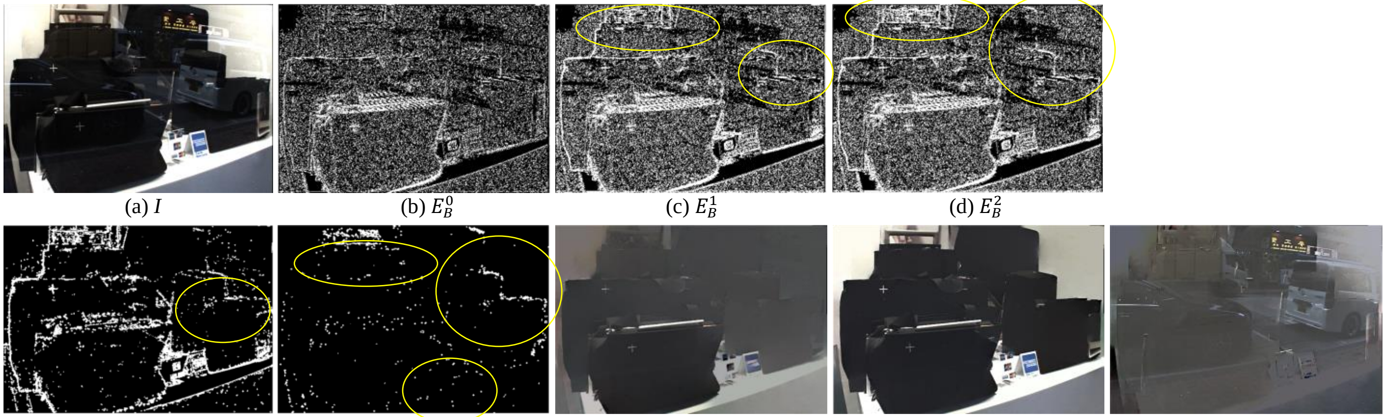
$$\begin{aligned} L^t = \arg\min_L F(L) = \sum_{p\epsilon\varphi^t} \Bigg(& U_p(L_p) \\ & + \lambda_L \sum_{q\epsilon SKNN(p)} V_{p,q}(L_p, L_q) \Bigg), \end{aligned} \tag{13}$$

$$U_p(L_p) = U_R^t(p)(1 - L_p) + U_B^t(p)L_p, \tag{14}$$

$$\begin{aligned} V_{p,q}(L_p, L_q) = \left(1 - N\left(\left\|\partial_{\bar{B}}^t(p) - \partial_{\bar{B}}^t(q)\right\|_1\right)\right) \\ \cdot \left[L_p \neq L_q\right], \end{aligned} \tag{15}$$

where λ_L is a constant, and the function $N\{x\}$ normalizes x to be between 0 to 1. The proposed energy function F in (13) is defined in such a way that its minimum corresponds to a good classification of the gradients in $I_{\bar{B}}^t$. L represents the label set. L_p denotes the label of the gradient at position p in set φ^t. It is set to 1 for the background gradient and 0 for the reflection gradient. The data term $U_p(L_p)$ penalizes the cost function if a wrong classification of L_p is made. More specifically, if the gradient of $I_{\bar{B}}^t$ at p belongs to the background but is incorrectly classified as a reflection (i.e., L_p is set to 0), $U_p(L_p)$ will have a large value since $U_R^t(p)$ is large in this case. On the other hand, if the gradient of $I_{\bar{B}}^t$at p belongs to the reflection but is incorrectly classified as background (i.e., L_p is set to 1), $U_p(L_p)$ will also have a large value since $U_B^t(p)$ is large in this case.

Similar to MRF, the data term U_p is supplemented with a smoothness term $V_{p,q}$ in (13), which measures the smoothness of the gradients in $I_{\bar{B}}^t$. It is observed that strong gradients, such as the edges of an object, are smooth in some orientation. Adjacent gradients with similar orientations likely belong to the same object in the same layer. Thus, the smoothness term in (13) is designed in such a way that it will be large and penalize the cost function F if neighboring gradients in $I_{\bar{B}}^t$ with similar orientations are assigned different labels. In (15), the function $[L_p \neq L_q] = 1$ if $L_p \neq L_q$; it is 0 otherwise. Thus, the term $V_{p,q}$ of two pixels p and q in $I_{\bar{B}}^t$ will be zero if they have the same label. Otherwise, $V_{p,q}$ is evaluated based on the 1-norm difference of the gradients $\partial_{\bar{B}}^t$. Note that $F(L)$ in (13) is evaluated by accumulating $V_{p,q}$ for all pixel pairs $\{p,q\}$ within the similarity-based KNN (*SKNN*) set of p, which is defined as the set of K nearest neighboring pixels (where K is chosen as 7) of p measured by the similarity in the gradient value and distance. Normally, all of the pixels within the *SKNN* set should have the same label due to the smoothness of the object gradients. If a pixel q within the set is wrongly classified, then the classification of p will still follow the majority in the set since $V_{p,q}$ is small. In a situation in which p is wrongly classified such that it is different from most of the others in the set, a large sum of $V_{p,q}$ will be generated. This approach penalizes the cost function and forces the label of p to change.


(a) I (b) E_B^0 (c) E_B^1 (d) E_B^2

(e) S^1 (f) S^2 (g) I_B^0 (h) I_B (i) I_R

Fig. 6. (a) The original image I. (b) The estimated initial gradient mask E_B^0. (c) The improved gradient mask E_B^1. (d) The improved gradient mask E_B^2. See the improved estimation (circled). (e) Mask S^1. (f) Mask S^2. (g) The estimated initial background layer I_B^0. (h) The final background layer I_B. (i) The final reflection layer I_R.

The optimization problem in (13) can be solved by the max-flow/min-cut method [35]. Finally, a mask based on L is generated as follows:

$$S^{t+1} = \rho\{L^t = 1\}, \qquad (16)$$

where $\rho\{x\}$ represents a morphological dilation operator with size 2×2 within the set φ^t. It is used since we assume that the neighboring gradients around the classified labels are also likely to belong to the same layer. Note that S^t can be considered to be a mask of the gradients that appear at the same positions as both I_B^t and the residue. It thus has included the gradients of the missing background components based on the argument discussed earlier in this section. Thus, using S^t, we update the initial gradient masks as follows:

$$\begin{aligned} E_B^t &= E_B^{t-1} \cup S^t \cup E_B^0 \cap (\sim E_R^0); \\ E_R^t &= E_{\bar{B}}^{t-1} \cap (\sim S^t) \cap (\sim E_B^0) \cup E_R^0 , \end{aligned} \qquad (17)$$

for $t > 0$. $E_{\bar{B}}^t$ is defined in (9). Recall that E_B^0 is estimated with a conservatively selected disparity threshold. Most of the gradient points that it covers belong to the background, although many of the background's gradient points can be missing. To enhance E_B^0, we first exclude those also covered by the reflection gradient mask E_R^0. Then, we add back those covered by S^t to E_B^{t-1} in each iteration, as shown in (17). With the improved estimation of I_B^t in each iteration, the estimation of S^t will also improve and in turn enhance the estimation of E_B^t. The design of E_R^t follows a similar philosophy. The new gradient masks now include the information of the missing components. They can be used to refine the background estimate as follows:

$$\begin{aligned} I_B^t = \arg\min_{I_B^t} J = \|D * I_B^t\|_1 + \|D * I_R^t\|_1 + \\ \lambda_1 \|E_B^t \cdot D * I_R^t\|_1 + \lambda_1 \|E_R^t \cdot D * I_B^t\|_1; \qquad (18) \\ \text{s.t. } I_R^t = I_{\bar{B}}^t = I - I_B^t; \text{ for } t > 0. \end{aligned}$$

In (18), λ_1 is a constant. Note that unlike the existing approach, there is only one optimization parameter I_B^t in (18) (we can find I_R^t by $I_R^t = I_{\bar{B}}^t = I - I_B^t$). This fact reduces the possibility that the optimization process falls into the wrong local minimum. Similar to (10), we use IRSL to minimize (18). We iteratively update the background layer until it converges (e.g., the change in the recovered I_B^t becomes very small). The whole algorithm is summarized below:

```
**Algorithm:**
Generate the disparity map using (2) to (6).
Obtain the initial estimates $E_B^0$, $E_R^0$, $I_B^0$ and $I_R^0$ using (7) to (10).
Set $t$ = 0.
**While not converge**
    $t = t + 1$;
    Compute $S^t$ using (12) to (16);
    Compute $E_B^t$, $E_R^t$ using (17);
    Obtain $I_B^t$, $I_R^t$ using (18);
**End**
**Output** $I_B = I_B^t$, $I_R = I_R^t$
```

Fig. 6 shows an example of the proposed algorithm at different stages of the operations. It can be seen in Fig. 6(g) that the initially estimated background has many components missing. The reason is that the initial gradient mask E_B^0 misses out on many strong gradients as shown in Fig. 6(b). With the help of S^1, as shown in Fig. 6(e), the updated gradient mask E_B^1 (Fig. 6(c)) starts to restore some of the missing components. It in turn improves the estimation of S^2 (Fig. 6(f)) and then E_B^2 (Fig. 6(d)), as seen in the circled regions. Note that while more and more missing background components are recovered in S^2 (see the upper two circles in Fig. 6(f)), we also notice that the mask covers fewer background gradient points (such as those in the lower circle in Fig. 6(f)). The reason is that with the improved estimation of I_B^1, there are fewer common gradient points with the residue of I_B^1, which means that they have been correctly recovered in I_B^1, and thus, S^2 does not need to include them. The final background image generated by the proposed algorithm is shown in Fig. 6(h). It shows a significant improvement over the initial guess. The resulting reflection image is shown in Fig. 6(i).

V. Comparisons and Evaluation

As mentioned above, reflection removal using light field images is a relatively new research topic; we find very few similar approaches in the literature. In this study, we evaluate the performance of the proposed algorithm by first comparing the layer separation using the disparity signs (LS-DS) method [17], which is relatively new and makes use of the LF camera to capture the multiple views of a scene in one shot. Hence, it can be applied to dynamic scenes in the proposed algorithm. However, LS-DS has a stringent requirement on the depth of the background and reflection layers, as well as the orientation of the camera. We will show in later sections how these restrictions affect the separation performance. In addition to [17], we compare the proposed algorithm with three other multiple-image-based reflection removal methods. They include the superimposed image decomposition (SID)[4], layer separation using SIFT flow (LS-SIFTF)[6], and layer separation using motion flow (LS-MF)[5] methods. All of these methods make use of the depth information of the scene to separate the background and reflection. Since these approaches capture the multiple views of a scene using a sequential approach, they can only be used in static scenes. In the comparisons, SID and SIFTF are implemented with the source code published by the authors. All free parameters are chosen in the same way as in their original programs. For LS-MF and LS-DS, their program code is not publicly available. We follow their papers and implement the algorithms by ourselves. For all of the free parameters in LS-MF and LS-DS, we follow as much as possible their papers. Only for the edge confidence threshold of LS-DS, we cannot find the exact way to determine its value in the paper. We set it to be 0.1 in the comparison, and it shows decent performance in general. With regard to the proposed algorithm, we set the free parameters λ, λ_L and λ_1 to 1, 1, and 1.5, respectively. In fact, we do not find that the final results are sensitive to their selection.

Note that since we use the Lytro Illum LF cameras to capture the required LF images, the images given by the proposed algorithm can only have the spatial resolution of 625 × 434 pixels, the same as the Lytro Illum. It is incomparable with those given by the non-LF multiple-image-based methods. However, the non-LF multiple-image-based methods also have problems in the temporal resolution, since the images must be taken sequentially. To make a fair comparison, we ignore the factors of spatial or temporal resolutions and focus only on the ability of different approaches in reflection removal, which is the main objective of this paper. In our comparison, we keep the spatial resolution of the testing images to be the same for all approaches. Additionally, all scenes are kept static such that the non-LF multiple-image-based methods will not be handicapped due to object motion. Under such conditions, we compare only

their reflection removal performance. Both qualitative and quantitative comparisons among these methods have been made. They are reported in the following sections.

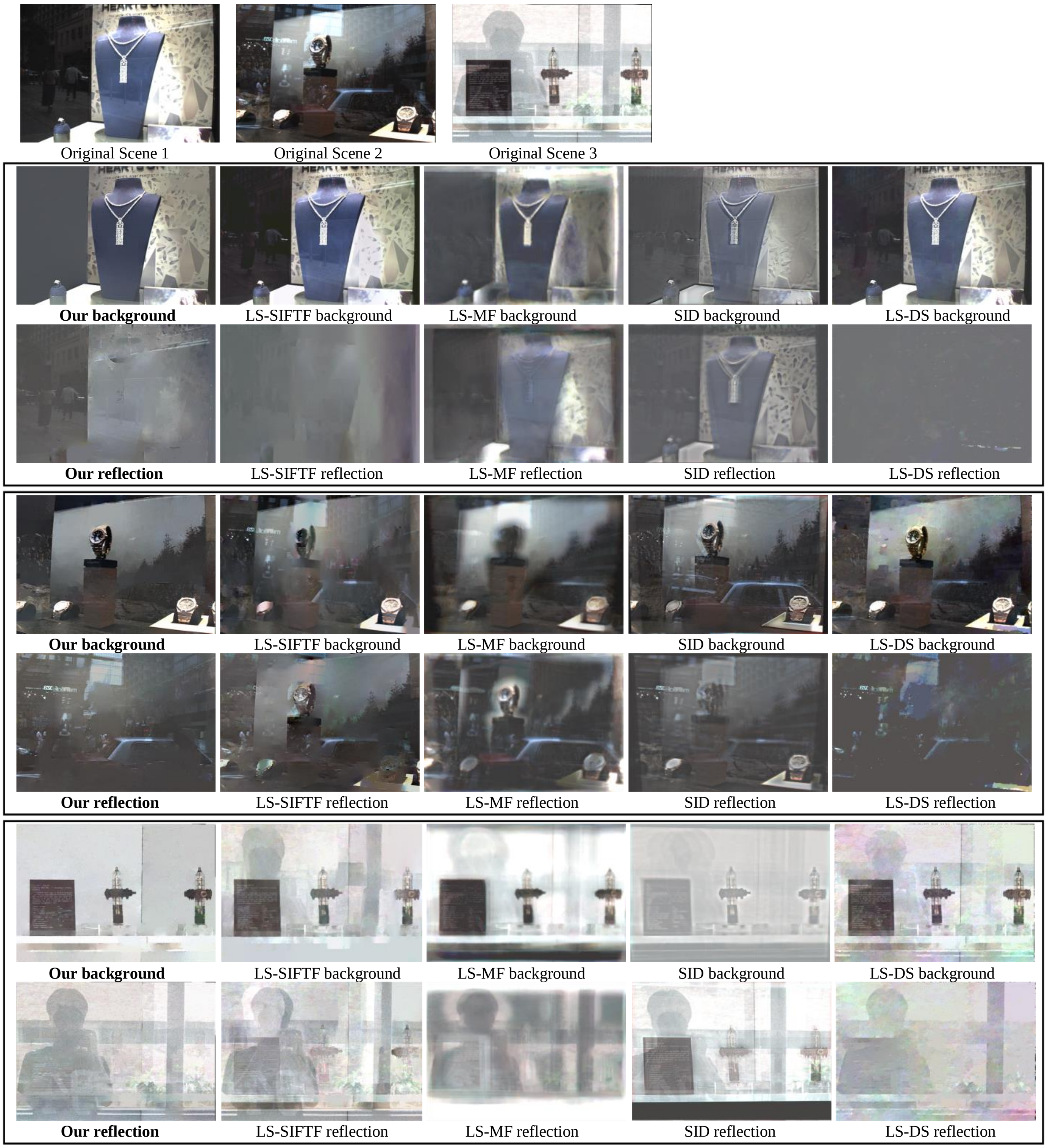
A. Qualitative Evaluation

For the qualitative evaluations, we compare visually the quality of the background and reflection images separated by different approaches. For testing the proposed algorithm and the method in [17], we make use of the Lytro Illum LF camera to obtain the LF images of a number of real-life scenes in which the background is superimposed on the reflection of an unwanted scene. The resolution of each view is 625×434 pixels, and we use the central 5×5 views to calculate the disparity. For the comparisons with the other multiple-image-based approaches [4, 5], we use the same LF camera to capture the same set of real-life scenes from an arbitrary 5 different viewing angles (because there is no specific requirement on the viewing angles for those multiple-image-based methods). Then, the central view of each LF image is collected to form the multiple-view images required by the methods [4, 5] and [6]. We show a few sets of comparison results in Fig. 7 and Fig. 8. Since they are all real-life scenes, there is no ground truth in all cases. However, from the contents in the separated background and reflection images, we can easily identify which approach performs the best.

As described in Section I and II, traditional methods all have their own requirements to the input images. It is difficult to ensure that they perform well for all images, particularly those taken from real scenes, since it is difficult to control the scene environment. As shown in Figs. 7 and 8, the recovered backgrounds tend to retain some residual reflections, and their reflection images also often contain background components. For example, it is noticed in the results of the method LS-SIFTF that it cannot separate the reflections that have strong gradients. This finding arises because the SIFT flow method used in LS-SIFTF can mistakenly register the strong reflection gradients as belonging to the background. As seen in its results for Scene 1, the reflection on the glass is largely retained in the resulting background image. It can be verified from the resulting reflection image, which contains almost no component of the reflection. For LS-MS, as mentioned in Sections I and II, the optimization process can easily fall into the wrong local minimum. We can find that the reconstructed background layers, which are the combination of all views, are blurred due to the inaccurate motion flows. For the SID method, it shows poor performance for scenes with non-planar background since it uses 2D homography to register the images. Moreover, the results of SID tend to be over-smooth because of the use of low rank decomposition with inaccurate registration. For LS-DS, it has a stringent requirement about the distance of the background or reflection layer. In many real-life scenes, such a requirement cannot be fully fulfilled. Additionally, it requires the normal axis of the LF camera to be aligned perpendicular to the scene. As shown in the images in Figs. 7 and 8, we often take pictures at an angle to the scene. This approach is about the style of the photography, and it is difficult to restrict. Since the scenes in Figs. 7 and 8 do not fully fulfill the requirements, the performance of LS-DS is only marginally satisfactory in most cases. As shown in the results for Scene 1, the reflection on the glass is largely retained in the resulting background image. Without the abovementioned limitations, the proposed algorithm can well reconstruct each layer and show the best performance in all cases. We also show a case with dynamic background in Fig. 9, where a television is showing a video behind a glass window. Since the TV screen is changing, the methods that require multiple shots of the scene from different angles cannot capture the same background and thus cannot be used in this case. Thus, we only test LS-DS and our method for this scene. Since the normal axis of the camera is not perpendicular to the scene, we can see that LS-DS leaves a large number of reflection residues on the background, while the proposed algorithm can give much better performance than LS-DS.

B. Quantitative evaluation

We have also conducted a quantitative comparison between the different approaches. The comparison with LS-DS and the proposed algorithm can be performed easily since both of them use the LF camera. In the comparison, we first use an LF camera to capture 20 LF images. Ten of them are selected as background, while the other ten are selected as reflection. They are manually added together to simulate the images that we need for the evaluation. Since the background is known, we can always measure the PSNR of the separated background with the true PSNR. To generate the images required for the evaluation of the multiple-image-based methods, we must have background and reflection images of different viewing angles for each scene. To do so, we do not take only one LF image for each scene, as mentioned above. Instead, we place the light field camera on a tripod and shift the camera at five fixed vertical heights to capture five LF images for every scene. Then, we use only the central view of each LF image such that for every scene, there are 5 images taken from 5 fixed vertical positions. Since all views of different scenes are taken at 5 fixed vertical heights, we can combine any two scenes together to simulate a background image with a reflection taken from 5 different viewing angles. In our experiments, we use the weightings 0.6 and 0.4 when combining the two images for generating the background and reflection, respectively. These images are then



Original Scene 1 | Original Scene 2 | Original Scene 3

Our background | LS-SIFTF background | LS-MF background | SID background | LS-DS background

Our reflection | LS-SIFTF reflection | LS-MF reflection | SID reflection | LS-DS reflection

Our background | LS-SIFTF background | LS-MF background | SID background | LS-DS background

Our reflection | LS-SIFTF reflection | LS-MF reflection | SID reflection | LS-DS reflection

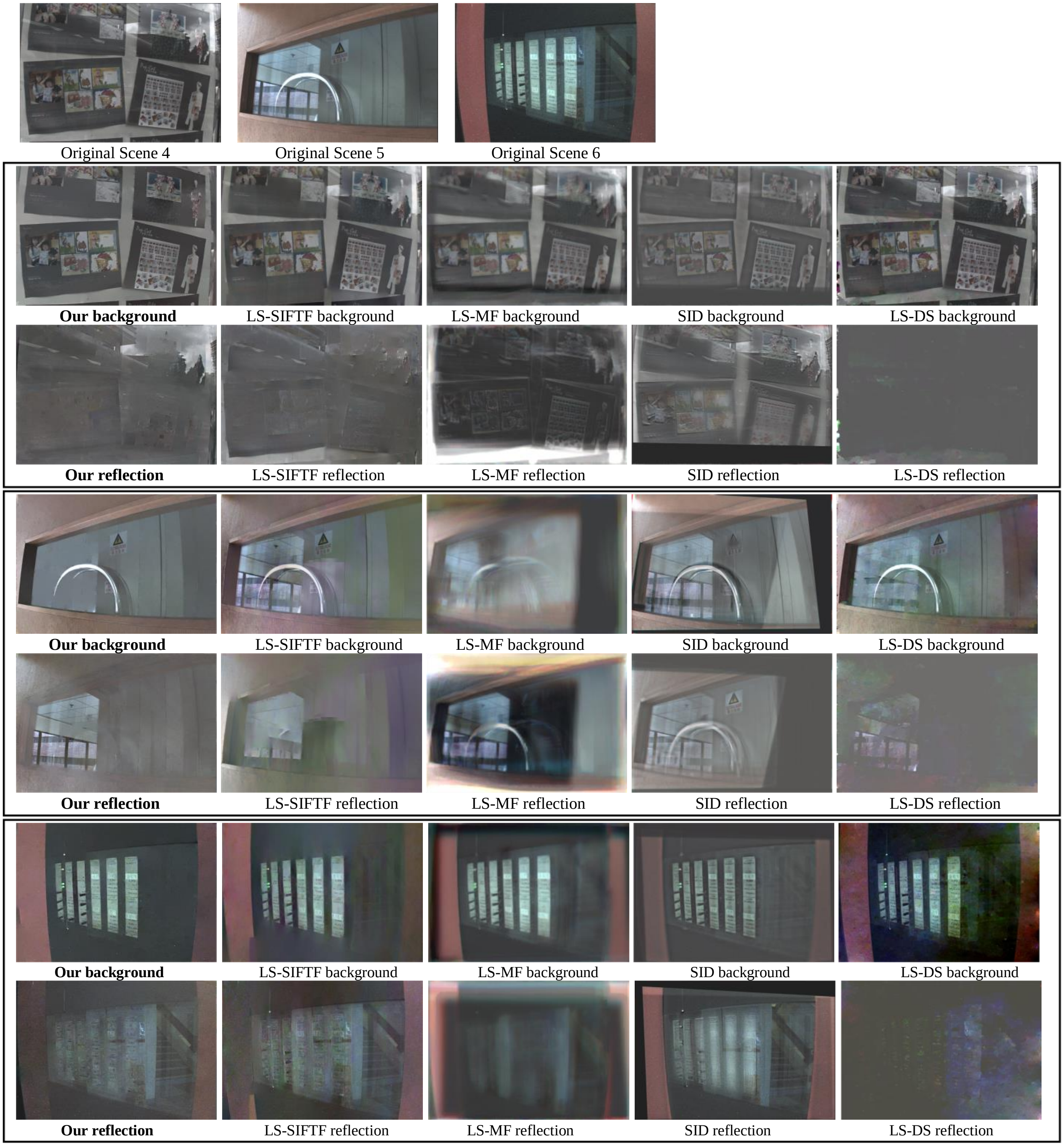
Our background | LS-SIFTF background | LS-MF background | SID background | LS-DS background

Our reflection | LS-SIFTF reflection | LS-MF reflection | SID reflection | LS-DS reflection

Fig. 7. Comparison results of scenes 1 to 3. For ease of visualization, the images are normalized by (11). Thus, for some images, the background plus reflection might not be equal to the original images. It can be seen that the proposed method shows robustness and better results compared to the other methods.

used in the evaluation of the multiple-image-based methods.

Since all separated images can contain biases, we adjust the bias of each separated image to achieve the maximum PSNR compared with the ground truths. Then, we compare the average maximum PSNR of the separated background and reflection images for all 10 scenes generated by all methods. The final results are shown in Table I. It can be seen that both LS-SIFTF and LS-DS can show relatively decent results compared to SID and LS-MF. The reason is that SID and LS-MF are established on or initialized with the planar background constraint while using homography. This constraint largely restricts their robustness compared to the other constraints of



Original Scene 4 Original Scene 5 Original Scene 6

Our background LS-SIFTF background LS-MF background SID background LS-DS background

Our reflection LS-SIFTF reflection LS-MF reflection SID reflection LS-DS reflection

Our background LS-SIFTF background LS-MF background SID background LS-DS background

Our reflection LS-SIFTF reflection LS-MF reflection SID reflection LS-DS reflection

Our background LS-SIFTF background LS-MF background SID background LS-DS background

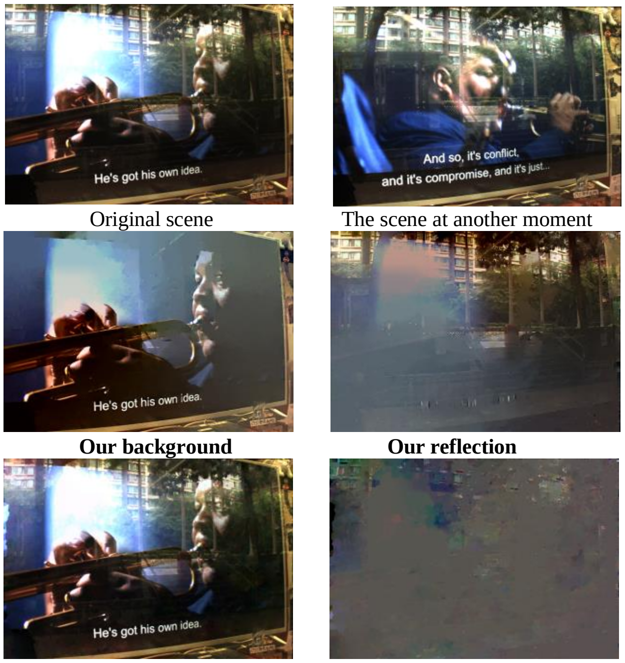
Our reflection LS-SIFTF reflection LS-MF reflection SID reflection LS-DS reflection

Fig. 8. Comparison results of scenes 4 to 6. For the ease of visualization, the images are normalized by (11). So for some images, the background plus reflection may not be equal to the original images. We can see that the proposed method shows robust and better results compared to other methods

LS-SIFTF and LS-DS. Although LS-MF attempts to refine the motions by optical flows, estimating too many variables simultaneously can cause it to be easily trapped into the wrong local minimums. In contrast, we can see that the proposed method without the abovementioned constraints outperforms all of the compared methods. The results are in line with the qualitative evaluation results.

C. *Computational Cost*

To evaluate the performance of the proposed algorithm in terms of the computational complexity, a comparison is made on the average execution times of the different methods for the six scenes shown in Figs. 7 and 8. All of the methods are executed in MATLAB, installed in a desktop computer with an Intel Core i7 7820X CPU and 64GB memory. From the results



Original scene The scene at another moment

Our background **Our reflection**

LS-DS background LS-DS reflection

Fig. 9. A dynamic scene case: a television behind a glass window. Since the scene in the television screen is changing, other methods that require multiple shots of the scene cannot work in this case. Therefore, only the results of LS-DS and the proposed method are shown. It can be seen that the proposed method gives much better performance than LS-DS.

TABLE I
THE AVERAGE PSNR VALUES OF THE SYNTHETIC INPUT IMAGES AND THE RESULTS OF EACH METHOD.

Method	First layer	Second layer
Synthetic input	13.0249	12.6774
LS-SIFTF	18.4999	18.9543
SID	15.2370	19.3243
LS-MF	16.5286	16.2398
LS-DS	18.6339	18.6433
Proposed	**21.9918**	**21.8188**

TABLE II.
THE AVERAGE EXECUTION TIME OF DIFFERENT METHODS TO GENERATE THE RESULTS IN FIGS. 7 AND 8.

Method	Average Time
LS-SIFTF	222.23 s
SID	218.67 s
LS-MF	623.94 s
LS-DS	61.15 s
Proposed	299.25 s

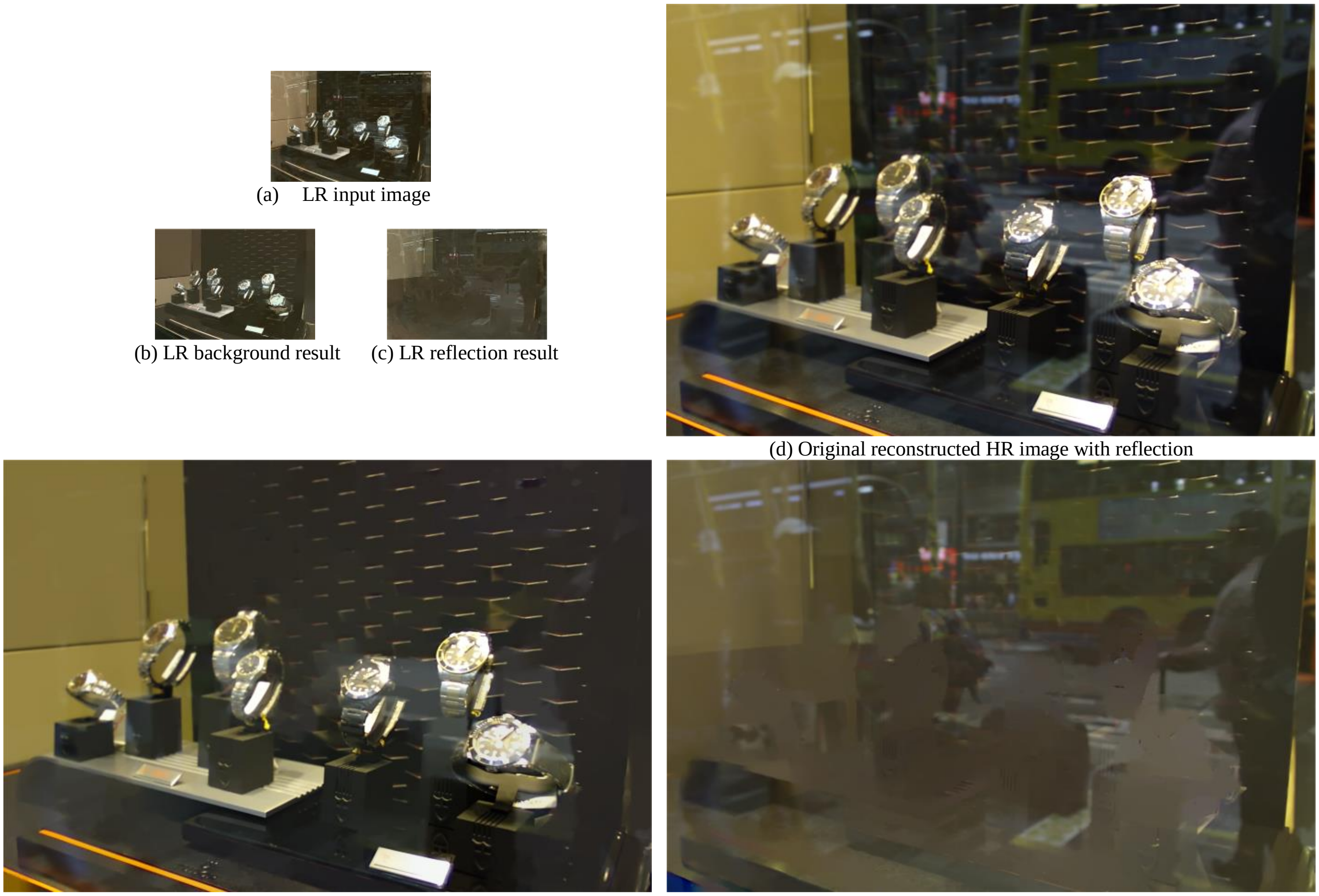
shown in Table II, it can be seen that LS-DS and LS-MF have the fastest and slowest speeds, respectively. The other three methods, LS-SIFTF, SID and the proposed method, are close in terms of the computational speed. We would like to emphasize that the current implementations of all of the comparison methods are only for proof-of-concept purposes. There is much room for optimizing the program code (and even the programming platform) before they can be used in practical applications , which furnishes one of the directions of our future work.

D. Possible extension to high-resolution (HR) images

One of the major deficiencies of the existing LF cameras is the spatial resolution of the captured light field. Since multiple views of a scene are captured on the same image sensor, the spatial resolution of each view is often much lower than the pixel resolution of the sensor. For example, Lytro Illum can only give LF images at 625×434 pixels. Current LF camera manufacturers solve the problem by using different super-resolution (SR) techniques [13, 36-38] to reconstruct the HR representation of the scene from the captured LR LF images. The qualities of the reconstructed HR images are comparable to those given by the traditional digital cameras. Since the reconstructed HR image has a close relationship with the LR LF images, it is possible to extend the proposed algorithm to assist in the reflection removal of the reconstructed HR image. To do so, we can first make use of the proposed algorithm to obtain the reflection removed LR background and reflection images. Then, we find their gradient masks E_B and E_R , respectively, and interpolate them to the resolution of the reconstructed HR image I^h. Let us denote the interpolated masks as E_B^h and E_R^h. Based on I^h, E_B^h and E_R^h, we can generate the reflection removed HR image by using the traditional gradient reconstruction method [6]. Fig. 10 shows an experimental result based on the abovementioned idea. In the experiment, we use Lytro's SR software to reconstruct an HR image I^h of resolution 2450×1634 pixels from the captured LR LF images (625×434 pixels). It can be seen in Fig. 10(d) that the image has a reflection image superimposed on the background. Based on the LR LF images, we use the proposed algorithm to generate the reflection removed background image and the reflection image (as shown in Fig. 10(b) and (c)). From these images, we obtain the gradient masks E_B and E_R and then bicubic interpolate them to obtain E_B^h and E_R^h, respectively. The resulting HR background and reflection images are obtained by applying I^h, E_B^h and E_R^h to a gradient reconstruction algorithm [6]. Fig. 10 (e) and (f) shows the final results. Because the SR software of Lytro automatically performs color rendering, the color style of the reconstructed HR image is different from the original LR LF images. However, we can still see that the reconstructed details are almost the same as the LR results. In addition, the reflection components in the reconstructed HR image are largely reduced. Note that the above experiment is only for proof-of-concept purposes. We believe that there is much room to further improve the performance. In addition, more experiments are needed to verify the method, which is another direction of our future work.

VI. CONCLUSIONS

In this paper, we proposed a novel algorithm for solving the reflection removal problem in photography based on light field imaging. One major improvement of the new algorithm is in its robustness, because it does not have the various restrictions on the scene or the camera orientation as in the existing approaches. In this paper, we first explored the behavior of the strong gradient points in the EPI of LF images when they are superimposed with reflection images. This approach provides



(a) LR input image

(b) LR background result

(c) LR reflection result

(d) Original reconstructed HR image with reflection

(e) HR reconstructed background

(f) HR reconstructed reflection

Fig. 10. A preliminary result of using the proposed algorithm to assist in HR image reflection removal. (a) The original LR LF image (central view); (b) and (c) results of using the proposed algorithm: background and reflection image, respectively; (d) original reconstructed HR image using the SR software of Lytro; (e) and (f) preliminary HR reflection removal results: background and reflection image, respectively. Because the HR images are too large to fit into the page size, we rescale the LR and HR images but keeping the size ratio to be 1:3.92 to illustrate their size difference.

theoretical support for using the light field imaging to estimate the disparities of the different layers of this type of image. We also proposed a general sandwich model to describe the disparity ranges of the components of the background and reflection layers. This step is the major part of how the proposed algorithm can be more versatile than the existing methods. Based on this model, we proposed a two-step strategy (initial conservative separation and component recovery) to well reconstruct the background layer in an iterative enhancement process. We have shown in the evaluation part that the proposed algorithm has a better and more robust performance compared to other methods. Since the light field camera can capture multiple views of a scene in a single shot, we believe the proposed algorithm can be further extended to the problem of video reflection removal, which is one of the directions that our group is working on. In addition, we have also demonstrated that the low resolution reflection removed image can be used to assist in the reflection removal of high-resolution images. We believe that further research in this direction will be fruitful.

Acknowledgments

This work is fully supported by the Hong Kong Polytechnic University under research grant RU9P.

References

[1] M. Born and E. Wolf, *Principles of optics: electromagnetic theory of propagation, interference and diffraction of light*: Elsevier, 2013.

[2] A. Cichocki and S.-i. Amari, *Adaptive blind signal and image processing: learning algorithms and applications*, 1st ed.: John Wiley & Sons, 2002.

[3] A. Levin and Y. Weiss, "User assisted separation of reflections from a single image using a sparsity prior," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 9, Sep 2007.

[4] X. Guo, X. Cao, and Y. Ma, "Robust separation of reflection from multiple images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 2187-2194.

[5] T. Xue, M. Rubinstein, C. Liu, and W. T. Freeman, "A computational approach for obstruction-free photography," *ACM Trans. Graph.*, vol. 34, no. 4, p. 79, Jul 2015.

[6] Y. Li and M. S. Brown, "Exploiting reflection change for automatic reflection removal," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 2432-2439.

[7] C. Liu, J. Yuen, and A. Torralba, "Sift flow: Dense correspondence across scenes and its applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 978-994, Aug 2011.

[8] M. Levoy and P. Hanrahan, "Light field rendering," in *Proc. Conf. Comput. Graph. Interactive Tech.*, 1996, pp. 31-42.

[9] R. Ng, M. Levoy, M. Brédif, G. Duval, M. Horowitz, and P. Hanrahan, "Light field photography with a hand-held plenoptic camera," *Comput. Sci. Tech. Rep.* , vol. 2, no. 11, pp. 1-11, Apr 2005.

[10] A. Isaksen, L. McMillan, and S. J. Gortler, "Dynamically reparameterized light fields," in *Proc. Conf. Comput. Graph. Interactive Tech.*, 2000, pp. 297-306.

[11] N. Joshi, S. Avidan, W. Matusik, and D. J. Kriegman, "Synthetic aperture tracking: tracking through occlusions," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2007, pp. 1-8.

[12] M. W. Tao, S. Hadap, J. Malik, and R. Ramamoorthi, "Depth from combining defocus and correspondence using light-field cameras," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 673-680.

[13] S. Wanner and B. Goldluecke, "Variational light field analysis for disparity estimation and super-resolution," *IEEE Trans. Pattern Anal. Mach. Intell.,* vol. 36, no. 3, pp. 606-619, Mar 2014.

[14] T.-C. Wang, A. A. Efros, and R. Ramamoorthi, "Occlusion-aware depth estimation using light-field cameras," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3487-3495.

[15] L. Jianqiao, L. Minlong, and L. Ze-Nian, "Continuous depth map reconstruction from light fields," *IEEE Trans. Image Process.,* vol. 24, no. 11, pp. 3257-3265, Nov 2015.

[16] T. Li and D. P. K. Lun, "Super-resolution imaging with occlusion removal using a camera array," in *Proc. IEEE Int. Sympos. Circuits Syst.*, 2016, pp. 2487-2490.

[17] Y. Ni, J. Chen, and L.-P. Chau, "Reflection Removal Based on Single Light Field Capture," in *Proc. IEEE Int. Sympos. Circuits Syst.*, 2017, pp. 1-4.

[18] R. C. Bolles, H. H. Baker, and D. H. Marimont, "Epipolar-plane image analysis: An approach to determining structure from motion," *Int. J. Comput. Vis.*, vol. 1, no. 1, pp. 7-55, Mar 1987.

[19] H. Farid and E. H. Adelson, "Separating reflections from images by use of independent component analysis," *JOSA A,* vol. 16, no. 9, pp. 2136-2145, Sep 1999.

[20] A. Agrawal, R. Raskar, S. K. Nayar, and Y. Li, "Removing photography artifacts using gradient projection and flash-exposure sampling," *ACM Trans. Graph.*, vol. 24, no. 3, pp. 828-835, Jul 2005.

[21] M. Wulkinson and J. Roerdink, Mathematical Morphology and Its Application to Signal and Image Processing. Berlin, Germany: Springer, 2009.

[22] Y. Li and M. S. Brown, "Single image layer separation using relative smoothness," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 2752-2759.

[23] Y. Shih, D. Krishnan, F. Durand, and W. T. Freeman, "Reflection removal using ghosting cues," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3193-3201.

[24] K. Gai, Z. Shi, and C. Zhang, "Blind separation of superimposed moving images using image statistics," *IEEE Trans. Pattern Anal. Mach. Intell.,* vol. 34, no. 1, pp. 19-32, Jan 2012.

[25] O. Johannsen, A. Sulc, and B. Goldluecke, "What sparse light field coding reveals about scene structure," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3262-3270.

[26] S. Wanner and B. Goldluecke, "Globally consistent depth labeling of 4D light fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 41-48.

[27] K. G. Derpanis, "The harris corner detector," *York Univ.*, 2004.

[28] J. Bigun, "Optimal orientation detection of linear symmetry," ed: Linköping University Electronic Press, 1987.

[29] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, "An efficient k-means clustering algorithm: Analysis and implementation," *IEEE Trans. Pattern Anal. Mach. Intell.,* vol. 24, no. 7, pp. 881-892, Jul 2002.

[30] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2005, pp. 886-893.

[31] A. Vedaldi and B. Fulkerson, "VLFeat: An open and portable library of computer vision algorithms," in *Proc. ACM Int. Conf. Multimedia*, 2010, pp. 1469-1472.

[32] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.,* vol. 32, no. 9, pp. 1627-1645, Sep 2010.

[33] S. Z. Li, "Markov random field models in computer vision," in *Proc. Eur. Conf. Comput. Vis.*, 1994, pp. 361-370.

[34] Q. Chen, D. Li, and C.-K. Tang, "KNN matting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 9, pp. 2175-2188, Sep 2013.

[35] Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision," *IEEE Trans. Pattern Anal. Mach. Intell.,* vol. 26, no. 9, pp. 1124-1137, Sep 2004.

[36] T. E. Bishop and P. Favaro, "The light field camera: Extended depth of field, aliasing, and superresolution," *IEEE Trans. Pattern Anal. Mach. Intell.,* vol. 34, no. 5, pp. 972-986, May 2012.

[37] R. A. Farrugia, C. Galea, and C. Guillemot, "Super resolution of light field images using linear subspace projection of patch-volumes," *IEEE J. Sel. Top. Sig. Proc.,* vol. 11, no. 7, pp. 1058-1071, Oct 2017.

[38] Z. Xu and E. Lam, "Light field superresolution reconstruction in computational photography," in *Proc. Conf. Sig. Recov. Synth.*, 2011, p. SMB3.