Depth from Semi-Calibrated Stereo and Defocus

Ting-Chun Wang
UC Berkeley
tcwang0509@berkeley.edu

Manohar Srikanth
Nokia Technologies
srimano@alum.mit.edu

Ravi Ramamoorthi
UC San Diego
ravir@cs.ucsd.edu

Abstract

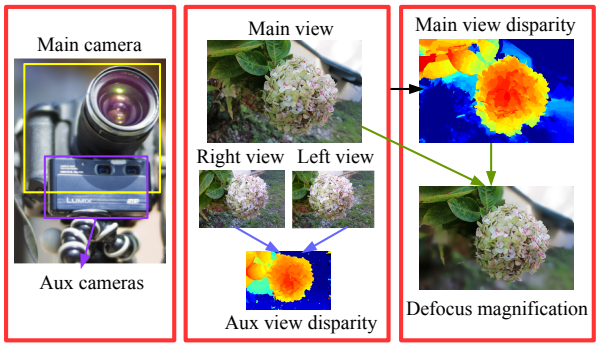
In this work, we propose a multi-camera system where we combine a main high-quality camera with two low-res auxiliary cameras. The auxiliary cameras are well calibrated and act as a passive depth sensor by generating disparity maps. The main camera has an interchangeable lens and can produce good quality images at high resolution. Our goal is, given the low-res depth map from the auxiliary cameras, generate a depth map from the viewpoint of the main camera. The advantage of our system, compared to other systems such as light-field cameras or RGBD sensors, is the ability to generate a high-resolution color image with a complete depth map, without sacrificing resolution and with minimal auxiliary hardware.

Since the main camera has an interchangeable lens, it cannot be calibrated beforehand, and directly applying stereo matching on it and either of the auxiliary cameras often leads to unsatisfactory results. Utilizing both the calibrated cameras at once, we propose a novel approach to better estimate the disparity map of the main camera. Then by combining the defocus cue of the main camera, the disparity map can be further improved. We demonstrate the performance of our algorithm on various scenes.

1. Introduction

Multi-camera systems have their advantages over traditional cameras, and recent commercial developments (e.g., Lytro [1] and Light [3]) have proven the market interest. In this paper, we show that by combining a high quality interchangeable-lens camera with auxiliary cameras that can recover depth (Fig. 1a), we can obtain many of the practical advantages of light field cameras [37]. When the photographer captures an image with the main camera, we also synchronously capture images from the auxiliary cameras (Fig. 1b). We then use these images to estimate the disparity map from the viewpoint of the main camera (Fig. 1c). This disparity map, along with the high-resolution main camera image, then enables a number of applications such as refocus magnification (Fig. 10) and parallax view generation (Fig. 11), while preserving image quality and resolution.

The reason we cannot just use two calibrated main cameras to recover depth is because the main camera has an interchangeable lens, so it is impossible to calibrate them beforehand. Besides, two auxiliary cameras are usually still much cheaper than a main camera. Finally, the auxiliary cameras are only loosely attached to the main camera to increase flexibility (one can easily detach them if we don't need depth maps). Therefore, calibration between main and auxiliary cameras is not possible.



(a) System setup (b) System input (c) System output

Figure 1: *System overview. (a) The system consists of a high-res uncalibrated main camera and two low-res calibrated auxiliary cameras. (b) The inputs contain the main camera image, the aux camera images, and the aux view disparity map. (c) Given these images, we generate the disparity map from the viewpoint of the main camera. We also show an example application of applying synthetic blur beyond the aperture baseline limit given the disparity map.*

To address this challenge, we are required to use uncalibrated rectification. However, the state-of-the-art results are still far from those generated by calibrated cameras. Therefore, instead of using two uncalibrated cameras, we use the (uncalibrated) main camera in conjunction with a calibrated pair of auxiliary cameras, e.g. the Panasonic Lumix DMC-3D1K [2], forming a semi-calibrated system. Our system thus contains a pair of calibrated low-res stereo cameras, and a high-res uncalibrated main camera. This is also similar to RGBD systems combining a color sensor and an active depth sensor (e.g. Kinect), where we want to transfer the depth to the viewpoint of the color sensor. However, RGBD cameras fail to perform well outdoors in sunlight, while our method is applicable indoors and outdoors. Besides, we can gain more rectification information from the color images of the calibrated camera pair, as we show in

Sec. 3. *To our knowledge, this is the first work that deals with semi-calibrated systems*.

We show that by decomposing the rectification into a 3-step approach, we can take advantage of the information from both the calibrated cameras at once, thus rectifying all three images simultaneously and more accurately. Then, by establishing the relationship between the three cameras, disparity estimation can be done through an optimization framework that ensures their disparity consistency.

In addition, when the main camera has defocus in the image, we also take that cue into account. The auxiliary cameras have small apertures, and can be considered as pinhole cameras; the main camera, on the other hand, has a large aperture and thus can generate very defocused images. By combining stereo and defocus cues, we show that we can estimate a higher quality disparity map.

In summary, our contributions are:

1) A method for combining high-res images with low-res auxiliary cameras for depth capture that provides both high quality imagery and depth maps.

2) A rectification scheme suitable for semi-calibrated systems, and an optimization framework to determine the disparities accordingly.

3) A method to combine defocus cues with stereo to help improve the disparity estimation process.

2. Related work

Related camera systems: Systems combining low and high resolution cameras have been proposed to perform depth map upsampling [13, 14, 17, 28, 31, 39, 48, 50] or color image upsampling [10, 11, 20, 42]. However, their goal is to increase the image resolution only, and the cameras in their system are usually assumed calibrated. In contrast, we are interested in generating a disparity map from the viewpoint of an uncalibrated camera, which is a very different and much harder problem, as we show in Sec. 5.

Some other work has been proposed to enable light-field capabilities to a consumer camera [35, 41]. However, for these methods the main camera image is deteriorated, which may not be acceptable. Moreover, the optical attachment proposed by Manakov *et al.* [35] is extremely large and not practical for consumer photography. In the case of Reddy *et al.* [41], the effective aperture of the camera is significantly reduced, thus cutting down the light input.

Uncalibrated rectification: Uncalibrated rectification dates back to Hartley *et al*. [23], where a linear and non-iterative method is given to reconstruct stereo by uncalibrated cameras. Loop and Zhang [33] decomposed each collineation into a specialized projective transform, a similarity transform and a shearing transform. Isgro and Trucco [26] proposed a method that avoids computation of the fundamental matrix. Mallon *et al*. [34] proposed a method that uniquely optimizes each transformation to minimize distortions. Fitzgibbon *et al*. [18] tried to learn the priors for calibrating families of stereo cameras A quasi-Euclidean method proposed by Fusiello *et al.* [19] aims at achieving a good approximation of the Euclidean epipolar rectification.

The previous methods all deal with two images. To rectify three images, Ayache *et al*. [6] proposed a technique to rectify image triplets from calibrated cameras. An *et al*. [4] proposed using the geometric camera model instead of the relative image orientation. For uncalibrated trinocular rectification, Sun [44] tried to rectify image triplets using the trilinear tensor, the projective invariants or fundamental matrices. Based on Sun's method, Heinrichs and Rodehorst [25] introduced a practical algorithm for various camera setups. Rectification methods extended to an arbitrary number of views with aligned camera centers have also been proposed [27, 38].

As can be seen, none of the previous methods tries to handle a system where only part of it is calibrated. Furthermore, all existing trinocular rectification methods either require the three camera centers to be collinear or to be non-collinear, and cannot handle both cases. In this work, we extend the quasi-Euclidean work proposed by Fusiello [19] to deal with three images, which generates much less distortion compared to the methods mentioned above.

Combining defocus with stereo: Since the main camera has a large aperture, we can also integrate depth from defocus (DFD) into the depth estimation framework. Most DFD work requires two or more images of the same scene [8, 21, 30, 40, 45, 46, 47]. Although we can construct an image pair using the auxiliary pinhole image and the main view defocused image, the two images are not registered, which makes DFD not reliable. However, we can still gain information using merely the main view image by single image DFD. For single image DFD, the work by Elder and Zucker [15] generated a sparse defocus map by applying first and second order derivatives to the input image. Bae *et al.* [7] extended this work and obtained a dense defocus map using interpolation. The deconvolution-based approach proposed by Levin *et al.* [29] can obtain both the depth map and the all-in-focus image, but relied on a coded aperture. Namboodiri and Chaudhuri [36] reversed a heat diffusion equation to estimate the defocus blur at edge locations. Zhuo and Sim [51] estimated the defocus map by re-blurring the input image using a known Gaussian blur kernel. Lin *et al.* [32] designed a sequence of filters to obtain an absolute depth map. In contrast, our method can generate the depth ratio between two sides of edges, which substantially enhances the depth map quality around occlusion edges during the final optimization.

3. Algorithm

As stated previously, our system consists of a main camera and two auxiliary cameras. The main camera has higher image quality, better resolution, and its optical parameters can be changed on the fly. The auxiliary cameras shoot

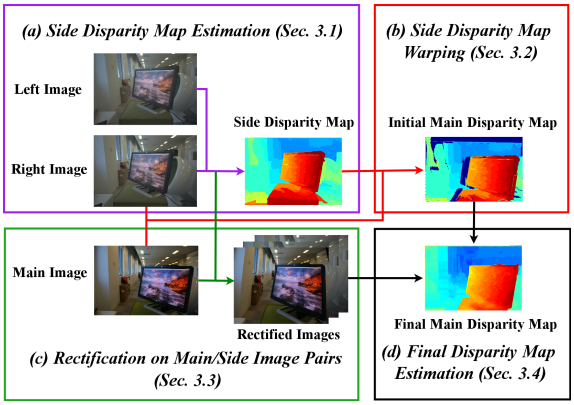


Figure 2: *Block diagram for our algorithm without defocus refinement. After we estimate a disparity map from the calibrated cameras, we warp it to the main view. This warped disparity map, along with the color images from main and aux cameras rectified using our method, are input to an optimization framework to generate the final disparity map. The refinement steps for defocus will be described in Sec. 4.*

smaller and lower quality images, and have fixed optical parameters. Our goal is to obtain the disparity map from the viewpoint of the main camera. To make statements easier, we will use the term *side cameras* to denote the auxiliary cameras, and use the terms *left camera* and *right camera* to differentiate between the two side cameras.

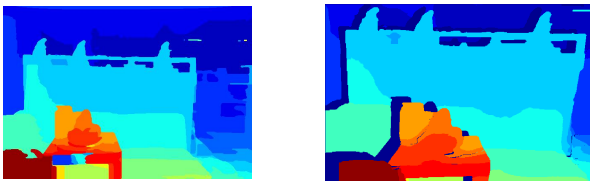
The overall algorithm, when the main camera has no defocus, can be divided into four steps. The refinement step for defocus will be described in Sec. 4. First, we calibrate the side cameras and estimate a disparity map from them (Sec. 3.1). Then this disparity map is warped to the main view to give us an initial disparity map (Sec. 3.2). Third, we perform uncalibrated rectification on the main image and the side images (Sec. 3.3), which is the main technical contribution of this section. Finally, the disparity map from the main view is estimated with the help of the warped side disparity map (Sec. 3.4). The block diagram of the overall algorithm is shown in Fig. 2.

3.1. Disparity estimation on side images

We first calibrate the two side cameras. To do this, we use the cameras to take multiple photos of a checkerboard at different orientations. After that, the Bouguet toolbox [12] is used to calibrate the cameras, so we can rectify the images. After rectification, we compute the disparity map using the non-local method proposed by Yang [49]. This is illustrated in Fig. 2a.

3.2. Side disparity map warping

Now we have the disparity map from the side point of view. To get a disparity map from the main point of view, one way is to warp the current side disparity map to the main view (Fig. 2b). We describe how we perform this here,



(a) Disparity from aux cameras (b) Warped disparity

Figure 3: *Side disparity map warping. After finding correspondences between the side view and the main view using NRDC, we estimate the projection matrix and warp the disparity map to the main view.*

and show how we utilize the warped disparity map to help estimate the final disparity map in Sec. 3.4.

First, we establish correspondences between the side view color image and the main view color image. Since the two views have very different color domains, direct feature matching will not lead to satisfactory results. Instead, we apply the non-rigid dense correspondence (NRDC) [22] on the two views which takes color transformation into account. Then we apply the epipolar constraint to rule out inconsistent matches. Finally, using the side camera coordinate as the world coordinate, we compute the projection matrix of the main camera, and warp the side view disparity to the main view. An example is shown in Fig. 3.

Note that for the pixels that are behind some other pixels, although they are not visible in the main view, we still know their corresponding positions in the side view. Thus, we can construct a correspondence map between the two images, where each pixel in the side view can be mapped to a position in the main view even if it is not visible there. This can be useful for applications such as inpainting, as demonstrated in our application section. Finally, there will definitely be holes in the warped disparity map due to occlusions. But that is acceptable since we only want to exploit the information that we can acquire from the calibrated disparity map, and the holes simply mean we need to rely on other sources, e.g. direct disparity estimation on the main/side image pair, as introduced next.

3.3. Rectification on main/side image pairs

Since the main camera is not calibrated, we are required to adopt uncalibrated rectification. Below we first give a review on previous methods, then introduce our method.

3.3.1 Background

Suppose we are given two images I_l and I_r to rectify. In other words, we want to find two homographies H_l and H_r such that when applied on the two images, their epipolar lines become entirely horizontal. After acquiring the corresponding feature points m_l and m_r in I_l and I_r, respectively, the estimation of the homographies can be formulated as

$$(H_r m_r^j)^T [u_1]_\times (H_l m_l^j) = 0 \tag{1}$$

where m_r^j and m_l^j are the jth corresponding feature points of I_r and I_l, respectively, and $[u_1]_\times$ is the skew-symmetric matrix associated with the cross-product by $u_1 = (1, 0, 0)$.

We adopt the quasi-Euclidean method by Fusiello *et al.* [19], where the homographies are decomposed as

$$H = K_n R K_o^{-1} \tag{2}$$

where K_n and K_o are the intrinsic matrices of the new and old cameras respectively, and R is the rotation matrix. Assuming negligible lens distortion, no skew and principal point at the image center, an intrinsic matrix is only dependent on the focal length. Therefore, the only parameters for each camera are the focal length and three rotation angles. Since rotation of one camera around its x-axis is redundant, we can further eliminate one degree of freedom, leaving only 7 free parameters.

3.3.2 Our method

The quasi-Euclidean method described above rectifies two images. To rectify three images, directly applying it will lead to solving 10 rotation angles and two focal lengths for the left-main and right-main image pairs (assuming common focal length between the left and the right cameras). In the following, we decompose the semi-calibrated rectification process into three steps, as summarized in Fig. 4. The general idea is to first bring the imaging planes of all the cameras to the same plane (step 1, 2), and then bring their x-axes to a common axis (step 3). By doing so, we show that we can reduce the rotation parameters to 6 angles.

Step 1 First, the calibrated side cameras are rotated by angles θ_{lx} and θ_{rx} respectively around the x-axis, so that their view direction is perpendicular to the plane the three cameras lie in (Fig. 4a). Since the side cameras are already rectified, $\theta_{lx} = \theta_{rx}$ and we will just use θ_{lx} for both of them. This angle always exists since the rotation axis, which is the line joining the two side camera centers, lies on the plane.

Step 2 Second, the main camera is rotated around the x-axis by an angle θ_{mx} and the y-axis by an angle θ_{my}, to bring its view direction parallel to the view direction of the side cameras (Fig. 4b).

Step 3 At this point, all three cameras have parallel view directions which are perpendicular to the plane they lie in, so we only need to rotate them around the z-axis to rectify them. We do this separately for the left-main and the right-main camera pair.

Left-main pair For the left and main camera pair, we rotate them by angles θ_{lz} and θ_{mz}, respectively (Fig. 4c,4d).

Right-main pair Since the left camera rotated by θ_{lz} and the main camera rotated by θ_{mz} are rectified, they have the same x-axis. Hence, when both are further rotated by $-\theta_{lz}$, they will still have parallel x-axes, which means the main camera rotated by $\theta_{mz} - \theta_{lz}$ will have parallel x-axis

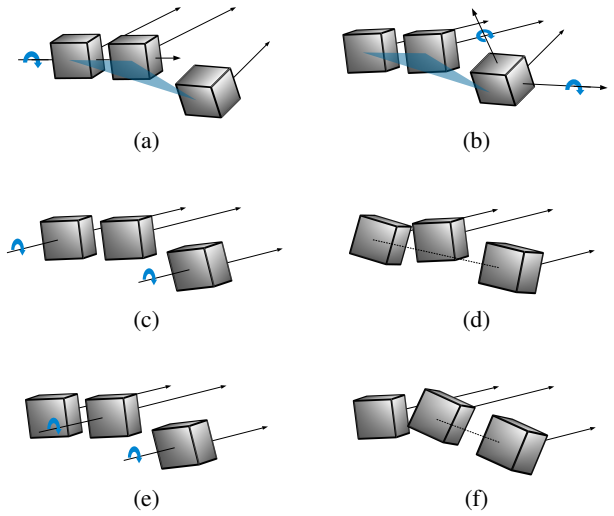


Figure 4: *Rectification on main/side image pair. The three boxes represent, from left to right, the left camera, the right camera, and the main camera. (a) The side cameras are rotated about their common x-axis to make their view axis perpendicular to the plane the three cameras lie in. (b) The main camera is rotated about the x- and y-axis to bring its view axis parallel to the view axes of side cameras. (c)-(d) The left camera and the main camera are rotated about their view axes to become rectified. (e)-(f) Similarly, the right and the main camera are rotated about their view axes.*

to the original left and right cameras. Then, to bring the main and the right cameras into a rectified setup, they must be further rotated by a common angle θ_{rz}. Therefore, for the right and main camera pair, we rotate them by angles θ_{rz} and $\theta_{mz} - \theta_{lz} + \theta_{rz}$, respectively (Fig. 4e,4f).

Thus, we only have 6 rotation parameters, namely $\theta_{lx}, \theta_{mx}, \theta_{my}, \theta_{lz}, \theta_{mz}$, and θ_{rz}. Adding the two parameters for focal length, this leaves us a total of 8 parameters. In implementation, we apply the standard Levenberg-Marquardt algorithm to solve for the parameters. This finishes our rectification process (Fig. 2c).

3.4. Final disparity map estimation

Now we have the warped disparity from the side view and the rectified main-side images, we formulate an optimization process to integrate them (Fig. 2d). Let I_m^l and I_l be the rectified main-left image pair, and I_m^r and I_r be the rectified main-right image pair. Then let d_{ml} be the disparity between I_m^l and I_l, d_{mr} be the disparity between I_m^r and I_r, and d_{lr} be the warped side view disparity as described in Sec. 3.2. Since the left, the right and the main camera are all in the same plane and have parallel viewing axes, d_{ml} and d_{mr} are proportional to each other (up to some known image rotation). It follows that $d_{ml} = C_{lr} d_{lr} = C_{mr} d_{mr}$, where C_{lr} and C_{mr} are two constants (the rotation notation is dropped for simplicity). As a result, d_{ml} can be estimated

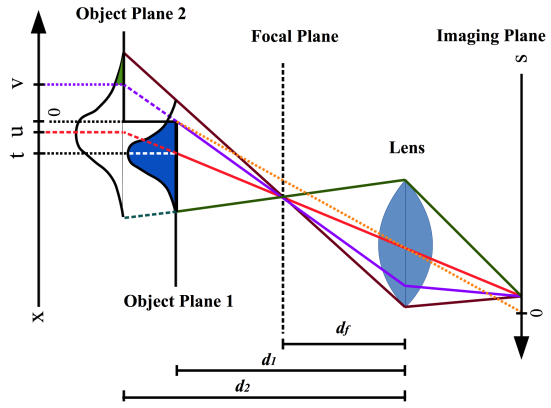


Figure 5: *Defocus model. We consider a 1D case, where a depth discontinuity is present in the scene. Due to occlusion, a pixel on the imaging plane is modeled as the sum of two partial Gaussians (the blue and green regions).*

as

$$\begin{aligned} d_{ml} = \arg\min_d \{ & \lambda_s \nabla d / \nabla I + \lambda_{lr} \| d - C_{lr} d_{lr} \|^2 \\ & + \lambda_{ml} \| I_m^l(x,y) - I_l(x+d,y) \|^2 \\ & + \lambda_{mr} \| I_m^r(x,y) - I_r(x+d/C_{mr},y) \|^2 \} \end{aligned} \tag{3}$$

where the first term applies the smoothness constraint, the second term ensures that d_{ml} matches d_{lr}, the third term is regular stereo matching between the main-left pair, and the final term ensures consistency to the main-right pair. All λ's are constant weights. In this way, we have a joint estimation incorporating all three disparities between the three cameras. Note that in practice, what we really want is the disparity map from the original view of the main camera; thus, the optimization is done in the following way: for each pixel in the original main image, we find its counterpart in the left/right images using the transformation we found in the rectification. We repeat this process for each disparity and compute their costs. This is similar to traditional stereo, only that we replace the horizontal shift by some other transformation function.

To obtain C_{lr} and C_{mr}, we first estimate an initial d_{ml} by computing the disparity between I_m^l and I_l. Then we use iteratively reweighted least squares (IRLS) to find a robust ratio C_{lr} that relates d_{ml} to d_{lr}. The same procedure is performed again to find C_{mr}.

4. Combining defocus cues

When the main image has defocus, we combine that cue into the disparity estimation. We consider two cues: defocus cue from the main image alone and the defocus cue from the main-side image pair, which are described as follows.

Defocus cue from main image Classical shape-from-defocus methods address the point spread function (PSF) as either a pillbox or a 2D Gaussian function [43]. In this work, we model the defocus as a Gaussian kernel. As shown by Bhasin *et al.* [9], the PSF around occlusions will be truncated and is no longer symmetric. Below we derive a simple constraint on disparities around occlusion edges based on the image gradients.

Pixel formation We first derive the formation of a pixel at an occlusion edge. Consider a simple 1D case where a depth discontinuity between two planes is present, as shown in Fig. 5. Similar to Favaro *et al.* [16] and Hasinoff *et al.* [24], we express the irradiance measured on the imaging plane using the reversed projection blurring model [5]. However, we make an assumption that around the edges, the radiance of the two planes is nearly constant. This dramatically reduces the computation. As can be seen, the intensity of that pixel is then modeled as the sum of two partial Gaussians of the two planes. Let the discontinuity be at origin $x = 0$, and let the chief ray (red line) hit plane 1 at $x = t$ and plane 2 at $x = u$. We first consider the case $t < 0$. Let the ray which hits plane 1 at $x = 0$ hit plane 2 at $x = v$. Due to similar triangles, $v - u = \frac{d_2 - d_f}{d_1 - d_f}(0 - t) = \frac{\sigma_2}{\sigma_1}(-t)$, where σ_1, σ_2 are the blur kernel sizes induced at plane 1 and 2, respectively. Let Φ be the cumulative Gaussian distribution. Then the sum of Gaussian on plane 1 is $\Phi_1(0 - t) = \Phi_1(-t)$, and the sum of Gaussian on plane 2 is $\Phi_2(u - v) = \Phi_2(\frac{\sigma_2}{\sigma_1} t)$. The case when $t > 0$ can be derived similarly. Finally, note that the world coordinate x and the image coordinate s are related by $x = s \cdot d / d_f$. Thus, the intensity of a pixel whose chief ray hits object plane at t is

$$I(s) \cong \begin{cases} I_1 \Phi_1(-\frac{\sigma_1}{\sigma_2}\frac{d_2}{d_f} s) + I_2 \Phi_2(\frac{d_2}{d_f} s), & s > 0 \\ I_1 \Phi_1(-\frac{d_1}{d_f} s) + I_2 \Phi_2(\frac{\sigma_2}{\sigma_1}\frac{d_1}{d_f} s), & s < 0 \end{cases} \tag{4}$$

where I_1 and I_2 are radiances at plane 1 and plane 2, respectively.

Blur kernel size Given the pixel formation, the blur kernel size can be obtained as follows. Let G denote the Gaussian function. Taking the derivative of I, and let $s \to 0$, we get

$$I'(s) = \begin{cases} -\frac{\sigma_1}{\sigma_2}\frac{d_2}{d_f} I_1 G_1(-\frac{\sigma_1}{\sigma_2}\frac{d_2}{d_f} s) + \frac{d_2}{d_f} I_2 G_2(\frac{d_2}{d_f} s), & s > 0 \\ -\frac{d_1}{d_f} I_1 G_1(-\frac{d_1}{d_f} s) + \frac{\sigma_2}{\sigma_1}\frac{d_1}{d_f} I_2 G_2(\frac{\sigma_2}{\sigma_1}\frac{d_1}{d_f} s), & s < 0 \end{cases} \tag{5}$$

$$I'(s)|_{s \to 0} = \begin{cases} \frac{d_2}{d_f}\frac{I_2 - I_1}{\sqrt{2\pi}\sigma_2}, & s \to 0^+ \\ \frac{d_1}{d_f}\frac{I_2 - I_1}{\sqrt{2\pi}\sigma_1}, & s \to 0^- \end{cases} \tag{6}$$

$$\frac{I'(0^+)}{I'(0^-)} = \frac{\sigma_1 / d_1}{\sigma_2 / d_2} \tag{7}$$

Therefore, the blur kernel sizes around an edge are related to the gradients around the edge.

Disparity constraint Finally, given the blur kernel size σ at a point, we can obtain the disparity d by [45]

$$\sigma = C_\sigma(d - d_f) \tag{8}$$

where C_σ is a constant and d_f is the disparity of the in-focus plane.

To use the above constraint, for the main image, we first apply the Canny edge detector to find the edges. Then we calculate d_f by taking the average of the in-focus regions in the disparity map computed before. Finally, the constraints on the disparities around the edges are modeled as

$$R = \frac{(d_1 - d_f)/d_1}{(d_2 - d_f)/d_2} = \frac{I'(0^+)}{I'(0^-)} \tag{9}$$

where d_1 and d_2 are disparities on the two sides of the edge, and $x = 0$ is the edge position.

Defocus cue from the main-side image pair We assume that the defocus image is the original sharp image convolved with a spatially variant Gaussian kernel. In other words, the defocus main image I^l_m, I^r_m and the sharp side images I_l, I_r are related by

$$I^l_m = I_l(x + d_{ml}, y) * G(\sigma(d_{ml})) \tag{10}$$

$$I^r_m = I_r(x + d_{mr}, y) * G(\sigma(d_{mr})) \tag{11}$$

where $G(\sigma)$ is a Gaussian with variance σ^2, and σ is a spatially variant function of the disparity value.

Combining stereo and both defocus cues Given the relationships we have above, we make two modifications to the previous optimization framework (Eq. (3)). *First*, we exploit the defocus cue from main-side image pair by blurring the side images according to the disparity before doing stereo matching. This can help improve the stereo matching process since the blurred side image will be more similar to the main image. *Second*, we apply the constraints on the disparities around edges we found using the defocus cue from the main image. This is beneficial because it explicitly models depth discontinuity at an edge, which is not handled by the defocus inference from the main-side image pair. As a result, by combining both cues, we are able to generate more accurate results along occlusion boundaries, while still preserving the dense property from matching between the main-side image pair.

We thus rewrite the optimization objective as

$$\begin{aligned} d_{ml} = \arg\min_d \{ & \lambda_s \nabla d / \nabla I + \lambda_{lr} \| d - C_{lr} d_{lr} \|^2 \\ & + \lambda_{ml} \| I^l_m(x, y) - I_l(x + d, y) * G(C_\sigma(d - d_f)) \|^2 \\ & + \lambda_{mr} \| I^r_m(x, y) - I_r(x + d/C_{mr}, y) * G(C_\sigma(C_{mr} d - d_f)) \|^2 \\ & + \lambda_r \sum_{\mathbf{l}_i} \sum_{p_1, p_2 \in \mathcal{N}(\mathbf{l}_i)} \| \frac{d(p_1) - d_f}{d(p_2) - d_f} - R(p) \|^2 \} \end{aligned} \tag{12}$$

where $\mathbf{l}_i$ is an edge found by the Canny edge detector, $\mathcal{N}$ is the 2 neighboring pixels lying on the normal of the edge, and $R(p)$ is the gradient ratio we computed from Eq. (9). The λ's are constant weights and are chosen as 5, 1, 1, 1, 2 in our experiment, respectively.

5. Results

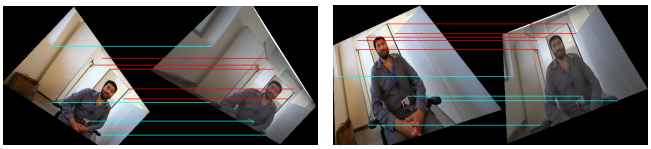
In this section, we evaluate our results of both the rectification and disparity map computation. We have captured data with a number of systems that have a main camera (Canon EOS 30D) and auxiliary cameras, such as the Panasonic Lumix DMC-3D1K Camera.

Rectification To perform a quantitative analysis, our rectification method is applied on 8 image sets and compared to the method by Fusiello *et al.* [19]. Since Fusiello *et al.* explicitly minimizes the errors on the extracted matching feature points, it might generate results that overfit the points. Thus, to make the comparisons fair, we divide the matching points into two sets, one for training and one for testing. For each tested algorithm, it first uses the training points to explicitly minimize the error. After that, the obtained homographies are used to calculate the errors on the testing points. The average Sampson errors [19] on the rectified images are summarized in Table 1. It can be seen that in general, our algorithm generates more accurate results. Moreover, when the average errors are close, e.g. image 4, the error variance generated by our algorithm is still smaller due to its robustness. An example visual result is shown in Fig. 6. By visual experiment, we found that our algorithm performs the best where few feature points are available. This is reasonable, because [19] explicitly minimizes the errors on the feature points, so it might overfit them while generating large errors on the less textured areas. On the other hand, our algorithm applies more constraints by ensuring consistency to both calibrated images, so the chances that it overfits the feature points are lower.

Disparity map We compare our disparity maps at different stages of the algorithm. We first show the results where defocus is not present in the main image in Fig. 7. It can be seen that for the warped disparity, the result is acceptable, but there will be many holes due to occlusions. For the result using our rectification, there will be no holes, but there will also be many errors due to incorrect rectification. By combining these two using the optimization in Sec. 3.4, the final disparity map can take advantage of both the warped disparity and the rectification. We also compare with results generated from direct depth upsampling [17] and simple uncalibrated rectification [19]. For [17], we use the warped side disparity maps as input. It can be seen that since conventional depth upsampling methods assume the depth image and the color image can be perfectly aligned, they will be misled by the errors in the disparity maps, e.g., the left side pillar in the first row. They also struggle to handle the large occlusions caused during the transfer; For instance, artifacts can be seen around the monitor in the first

#	Avg [19]	Avg (Ours)	Var [19]	Var (Ours)
1	1.3795	0.7433	1.8769	0.3684
2	1.8839	0.9083	1.9613	0.5396
3	1.4637	0.8586	1.5533	0.4481
4	0.9839	0.8865	1.2459	0.2784
5	0.9504	0.7159	1.1607	0.3468
6	3.8306	2.3171	2.9373	1.9889
7	1.3899	0.7596	1.5429	0.2869
8	3.8752	2.2736	4.4571	0.5109

Table 1: *Rectification error comparison. It can be seen that in general, our algorithm generates more accurate results. Moreover, when the average errors are close, e.g. image 4, the error variance generated by our algorithm is still smaller due to its robustness.*



(a) Fusiello *et al.* [19] (b) Our method

Figure 6: *An example rectification result on main/side image pairs. The cyan lines show where both algorithms perform well; the red lines show where the method by Fusiello et al. fails, while ours still generates reasonable results.*

row and around the bag in the second row. Finally, all results are much better than simply doing uncalibrated rectification [19].

Next, we compare results where defocus is present in the main image, using methods with and without our defocus refinement. We also compare with the results obtained by depth upsampling [17], Lytro Illum and Kinect version 2. For each scene, we take pictures using our setup as well as using the Lytro camera and Kinect, as shown in Fig. 8. We can see that by utilizing the defocus cue, we are able to generate much sharper boundaries along occlusion edges. We also get more accurate results on the background since its blur is taken into account now. Again, our results are superior than simple depth upsampling, since the imperfect calibration registers wrong depths that [17] uses as seeds. Moreover, we generate results which are much less noisy than the results of Lytro Illum. Using Kinect as ground truth, we can see that our results are very similar to its results, while we only use passive sensors. Besides, our disparity map has a higher resolution compared to the result of Kinect, so when zoomed in, there will be fewer artifacts, as shown in Fig. 9. For quantitative comparisons, we provide the depth RMSE using Kinect as ground truth in Table 2. We show the average RMSE of all indoor scenes, and compare with results by Ferstl et al. [17], Zhuo et al. [51], and Lytro Illum. We also provide results when only our stereo cue or defocus cue is used. The best result is achieved when combining both cues. Note that this comparison cannot capture the full benefit of incorporating defocus cues, since the result of Kinect does not have sharp boundaries around depth discontinuities.

Ferstl et al.	Zhuo et al.	Lytro Illum
0.0605	0.0727	0.0655
Ours (w/o defocus)	Our defocus only	Our final
0.0568	0.0651	***0.0558***

Table 2: *Depth RMSE using Kinect as ground truth.*

6. Applications

Defocus magnification Given the partially defocused main image and the disparity map, we can blur the defocus region even more to simulate the shallow depth-of-field of a larger aperture, as demonstrated by Bae *et al.* [7]. The result compared to [7] is shown in Fig. 10. Note that their result incorrectly blurs the cookie can in the foreground, while leaving most of the background (e.g. lights on the ceiling) unblurred. Our algorithm, by combining stereo and defocus cues, generates a more accurate depth map which then gives a more realistic result.

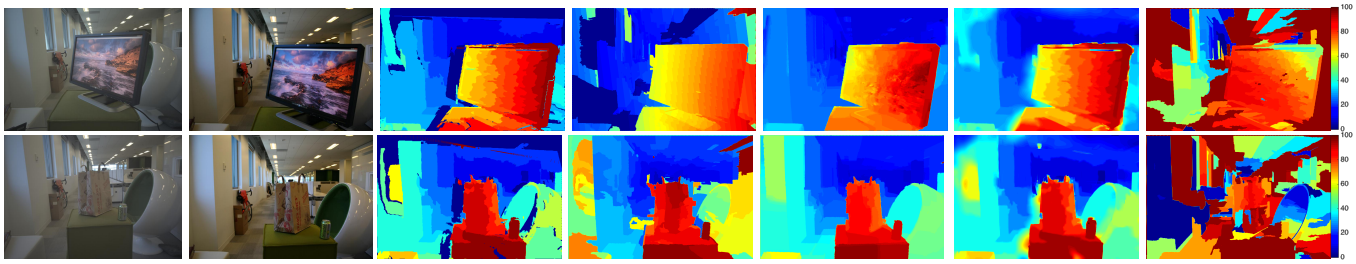
Parallax view generation Given the high-quality disparity map, we are able to generate realistic parallax views of the main image. However, generating parallax using only the disparity map will result in holes due to occlusion. To resolve this, we take advantage of the correspondence map introduced in Sec. 3.2. Once we know which pixel the hole corresponds to in the side views (assuming it exists), we can "borrow" that pixel to fill in the hole. We also use Poisson blending to compensate for the color changes. This is illustrated in Fig. 11.

7. Conclusion

In this paper, we propose a multi-camera system that includes an uncalibrated main camera and two calibrated auxiliary cameras. Previous works consider either entirely calibrated or uncalibrated systems. However, by exploiting the information from the calibrated cameras, we show that we can improve the disparity estimation of the main camera in two-fold. *First*, by decomposing the rectification process into three steps, we can rectify all three images at once, thus ensuring their consistency. An optimization framework to determine the disparity is also developed. *Second*, when defocus is present in the main image, we also take that into account. Defocus from a single image helps resolve the disparity ambiguities around edges, while disparity from an image pair can lead to a dense defocus map. By combining cues from both stereo and defocus, the disparity map is further improved. Utilizing this disparity map, we show we are able to achieve many of the applications of light field cameras, while still preserving high image quality.

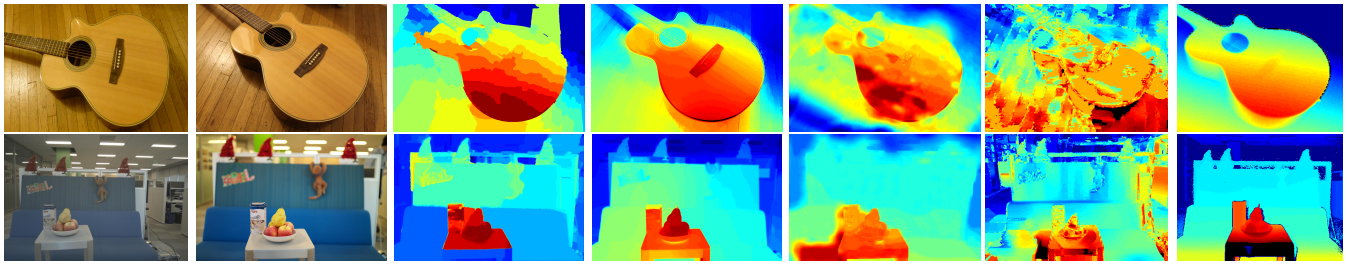
Acknowledgement

We thank Timo Ahonen for his helpful advice, the support and funding from Nokia, the UC San Diego Center for Visual Computing, and a Berkeley Fellowship.



(a) Side view input (b) Main view input (c) Warped from side view (Sec. 3.2) (d) Our rectification (Sec. 3.3) (e) Our method (Sec. 3.2-3.4) (f) Ferstl *et al.* [17] (g) Fusiello *et al.* [19]

Figure 7: *Disparity map results. (c) The disparity map warped from side view. (d) The disparity map estimated on images using the rectification method in Sec. 3.3. (e) The disparity map obtained using the optimization process in Sec. 3.4. (f) The disparity map obtained using depth upsampling [17] on warped side view disparity map. Some artifacts can be seen, e.g., on the left side pillar and around the monitor in the first row, and around the bag in the second row. (g) The disparity map estimated on images rectified by Fusiello et al.*

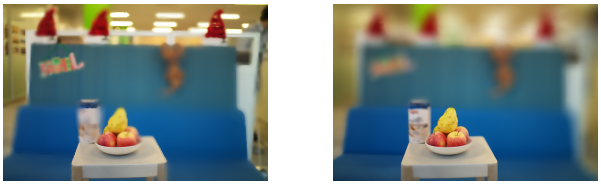


(a) Side view input (b) Main view input (c) Ours (w/o defocus) (d) Ours (w/ defocus) (e) Ferstl *et al.* [17] (f) Lytro Illum (g) Kinect

Figure 8: *Disparity map comparisons. (c)(d) By exploiting the defocus cues, our method generates more accurate results on the background (e.g. the sofa), and sharper boundaries around the foreground objects. (e) The disparity map obtained using depth upsampling [17]. It can be seen that since the side view disparity map and the main camera calibrations are not perfect, some errors will be registered to the main view, e.g. the floor in the first row and around the table in the second row. (f) The depth map from Lytro Illum. It can be seen that it is quite noisy, incorrectly labels some background as foreground, and does not produce as sharp edges as our method. (g) The depth map from Kinect. We can see that our result is very similar to Kinect, while we only use passive sensors.*

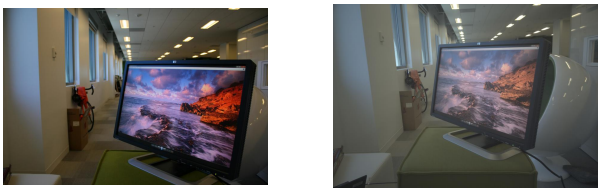
(a) Ours (b) Kinect

Figure 9: *Close-up comparison between our result and Kinect. It can be seen that our result, due to the higher resolution, has fewer artifacts.*

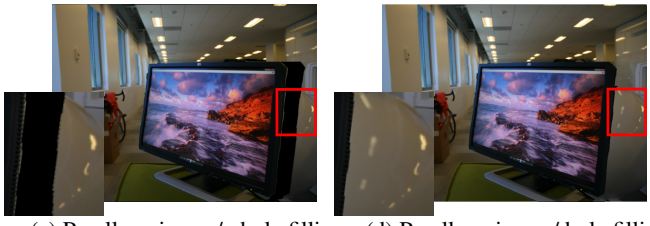


(a) Bae *et al.* [7] (b) Our method

Figure 10: *Defocus magnification applied on Fig. 8b. Note that in (a), the cookie can in the foreground is blurred, while most of the background is not blurred.*



(a) Original main view (b) Original left view



(c) Parallax view w/o hole filling (d) Parallax view w/ hole filling

Figure 11: *Parallax view generation. Note that (d) realistically reproduces the view behind the monitor from the side view, which is accomplished from the correspondence map introduced in the text, which contains more information than the disparity map.*

References

[1] Lytro redefines photography with light field cameras. Press release, Jun 2011. http://www.lytro.com. 1

[2] Panasonic LUMIX DMC-3D1K digital camera. http://shop.panasonic.com/support-only/DMC-3D1K.html/. 1

[3] Your smartphone camera could be a 52-megapixel beast by next year. Press release, Apr 2015. https://light.co. 1

[4] L. An, Y. Jia, J. Wang, X. Zhang, and M. Li. An efficient rectification method for trinocular stereovision. In *Proceedings of the International Conference on Pattern Recognition (ICPR)*, 2004. 2

[5] N. Asada, H. Fujiwara, and T. Matsuyama. Analysis of photometric properties of occluding edges by the reversed projection blurring model. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 20(2):155–167, 1998. 5

[6] N. Ayache and C. Hansen. Rectification of images for binocular and trinocular stereovision. In *Proceedings of the International Conference on Pattern Recognition (ICPR)*, 1988. 2

[7] S. Bae and F. Durand. Defocus magnification. *Computer Graphics Forum*, 26(3):571–579, 2007. 2, 7, 8

[8] R. Ben-Ari. A unified approach for registration and depth in depth from defocus. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 36(6):1041–1055, 2014. 2

[9] S. S. Bhasin and S. Chaudhuri. Depth from defocus in presence of partial self occlusion. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2001. 5

[10] P. Bhat, C. L. Zitnick, N. Snavely, A. Agarwala, M. Agrawala, M. Cohen, B. Curless, and S. B. Kang. Using photographs to enhance videos of a static scene. In *Proceedings of the Eurographics Conference on Rendering Techniques*, 2007. 2

[11] V. Boominathan, K. Mitra, and A. Veeraraghavan. Improving resolution and depth-of-field of light field cameras using a hybrid imaging system. In *Proceedings of the IEEE International Conference on Computational Photography (ICCP)*, 2014. 2

[12] J.-Y. Bouguet. Camera calibration toolbox for Matlab, Dec. 2013. 3

[13] J. Diebel and S. Thrun. An application of Markov random fields to range sensing. In *Proceedings of the International Conference on Neural Information Processing Systems (NIPS)*, 2005. 2

[14] J. Dolson, J. Baek, C. Plagemann, and S. Thrun. Upsampling range data in dynamic environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010. 2

[15] J. H. Elder and S. W. Zucker. Local scale control for edge detection and blur estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 20(7):699–716, 1998. 2

[16] P. Favaro and S. Soatto. Seeing beyond occlusions (and other marvels of a finite lens aperture). In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2003. 5

[17] D. Ferstl, C. Reinbacher, R. Ranftl, M. Rüther, and H. Bischof. Image guided depth upsampling using anisotropic total generalized variation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2013. 2, 6, 7, 8

[18] A. W. Fitzgibbon, D. P. Robertson, A. Criminisi, S. Ramalingam, and A. Blake. Learning priors for calibrating families of stereo cameras. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2007. 2

[19] A. Fusiello and L. Irsara. Quasi-euclidean epipolar rectification of uncalibrated images. *Machine Vision and Applications*, 22(4):663–670, 2011. 2, 4, 6, 7, 8

[20] D. C. Garcia, C. Dorea, and R. L. De Queiroz. Super resolution for multiview images using depth information. *IEEE Transactions on Circuits and Systems for Video Technology*, 22(9):1249–1256, 2012. 2

[21] I. Gheta, C. Frese, M. Heizmann, and J. Beyerer. A new approach for estimating depth by fusing stereo and defocus information. In *GI Jahrestagung (1)*, pages 26–31, 2007. 2

[22] Y. HaCohen, E. Shechtman, D. B. Goldman, and D. Lischinski. Non-rigid dense correspondence with applications for image enhancement. *ACM Transactions on Graphics (TOG)*, 30(4):70, 2011. 3

[23] R. Hartley, R. Gupta, and T. Chang. Stereo from uncalibrated cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1992. 2

[24] S. W. Hasinoff and K. N. Kutulakos. A layer-based restoration framework for variable-aperture photography. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2007. 5

[25] M. Heinrichs and V. Rodehorst. Trinocular rectification for various camera setups. In *Symposium of ISPRS Commission III-Photogrammetric Computer Vision (PCV)*, 2006. 2

[26] F. Isgro and E. Trucco. Projective rectification without epipolar geometry. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1999. 2

[27] Y.-S. Kang, C. Lee, and Y.-S. Ho. An efficient rectification algorithm for multi-view images in parallel camera array. In *Proceedings of the 3DTV Conference*, 2008. 2

[28] Y. M. Kim, C. Theobalt, J. Diebel, J. Kosecka, B. Miscusik, and S. Thrun. Multi-view image and ToF sensor fusion for dense 3D reconstruction. In *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2009. 2

[29] A. Levin, R. Fergus, F. Durand, and W. T. Freeman. Image and depth from a conventional camera with a coded aperture. In *ACM Transactions on Graphics (TOG)*, volume 26, page 70, 2007. 2

[30] C. Li, S. Su, Y. Matsushita, K. Zhou, and S. Lin. Bayesian depth-from-defocus with shading constraints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013. 2

[31] F. Li, J. Yu, and J. Chai. A hybrid camera for motion deblurring and depth map super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008. 2

[32] J. Lin, X. Ji, W. Xu, and Q. Dai. Absolute depth estimation from a single defocused image. *IEEE Transactions on Image Processing*, 22(11):4545–4550, 2013. 2

[33] C. Loop and Z. Zhang. Computing rectifying homographies for stereo vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1999. 2

[34] J. Mallon and P. F. Whelan. Projective rectification from the fundamental matrix. *Image and Vision Computing*, 23(7):643–650, 2005. 2

[35] A. Manakov, J. F. Restrepo, O. Klehm, R. Hegedüs, E. Eisemann, H.-P. Seidel, and I. Ihrke. A reconfigurable camera add-on for high dynamic range, multispectral, polarization, and light-field imaging. *ACM Transactions on Graphics (TOG)*, 32(4):47, 2013. 2

[36] V. P. Namboodiri and S. Chaudhuri. Recovery of relative depth from a single observation using an uncalibrated (real-aperture) camera. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008. 2

[37] R. Ng. *Digital light field photography*. PhD thesis, Stanford University, 2006. 1

[38] V. Nozick. Multiple view image rectification. In *IEEE International Symposium on Access Spaces (ISAS)*, 2011. 2

[39] J. Park, H. Kim, Y.-W. Tai, M. S. Brown, and I. Kweon. High quality depth map upsampling for 3D-ToF cameras. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2011. 2

[40] A. Rajagopalan, S. Chaudhuri, and U. Mudenagudi. Depth estimation and image restoration using defocused stereo pairs. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 26(11):1521–1525, 2004. 2

[41] D. Reddy, J. Bai, and R. Ramamoorthi. External mask based depth and light field camera. In *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2013. 2

[42] H. S. Sawhney, Y. Guo, K. Hanna, R. Kumar, S. Adkins, and S. Zhou. Hybrid stereo camera: an IBR approach for synthesis of very high resolution stereoscopic image sequences. In *Proceedings of Siggraph*, 2001. 2

[43] M. Subbarao and G. Surya. Depth from defocus: a spatial domain approach. *International Journal of Computer Vision (IJCV)*, 13(3):271–294, 1994. 5

[44] C. Sun. Uncalibrated three-view image rectification. *Image and Vision Computing*, 21(3):259–269, 2003. 2

[45] Y. Takeda, S. Hiura, and K. Sato. Fusing depth from defocus and stereo with coded apertures. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013. 2, 6

[46] M. W. Tao, S. Hadap, J. Malik, and R. Ramamoorthi. Depth from combining defocus and correspondence using light-field cameras. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2013. 2

[47] V. Vaish, M. Levoy, R. Szeliski, C. L. Zitnick, and S. B. Kang. Reconstructing occluded surfaces using synthetic apertures: Stereo, focus and robust measures. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006. 2

[48] J. Yang, X. Ye, K. Li, and C. Hou. Depth recovery using an adaptive color-guided auto-regressive model. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 2012. 2

[49] Q. Yang. A non-local cost aggregation method for stereo matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 3

[50] Q. Yang, R. Yang, J. Davis, and D. Nistér. Spatial-depth super resolution for range images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007. 2

[51] S. Zhuo and T. Sim. Defocus map estimation from a single image. *Pattern Recognition*, 44(9):1852–1858, 2011. 2, 7