applied sciences

MDPI

Article

Iterative Refinement of Uniformly Focused Image Set for Accurate Depth from Focus

Sherzod Salokhiddinov and Seungkyu Lee *

Department of Computer Science & Engineering, Kyung Hee University, Yongin-si 446-701, Korea; sherzod@khu.ac.kr

* Correspondence: seungkyu@khu.ac.kr

Received: 31 October 2020; Accepted: 26 November 2020; Published: 28 November 2020

check for updates

Abstract: Estimating the 3D shape of a scene from differently focused set of images has been a practical approach for 3D reconstruction with color cameras. However, reconstructed depth with existing depth from focus (DFF) methods still suffer from poor quality with textureless and object boundary regions. In this paper, we propose an improved depth estimation based on depth from focus iteratively refining 3D shape from uniformly focused image set (UFIS). We investigated the appearance changes in spatial and frequency domains in iterative manner. In order to achieve sub-frame accuracy in depth estimation, optimal location of focused frame in DFF is estimated by fitting a polynomial curve on the dissimilarity measurements. In order to avoid wrong depth values on texture-less regions we propose to build a confidence map and use it to identify erroneous depth estimations. We evaluated our method on public and our own datasets obtained from different types of devices, such as smartphones, medical, and normal color cameras. Quantitative and qualitative evaluations on various test image sets show promising performance of the proposed method in depth estimation.

Keywords: depth estimation; depth from focus; 3D reconstruction; uniformly focused image set

1. Introduction

Estimating three dimensional shape of a scene from color image is a challenging task [1]. Without any prior knowledge on the scene, it is an ill-posed problem to recover three dimensional shape of objects using single color camera. Many researchers have proposed diverse approaches for 3D depth acquisition. For example, in Reference [2,3], the authors estimate depth information based on a coded aperture framework. However, it requires hardware modification which cannot be applied to off-the-shelf camera, such as smart phone camera. Depth from focus (DFF), or shape from focus (SFF), is a technique for depth estimation from a set of image frames having continuously changing focus amount that are taken at the same location and viewpoint. By identifying best focused frame of each pixel, depth of each pixel can be estimated. In DFF, three factors determine the quality of reconstructed depth: lens aperture, focal length, and focusing distance. While many computer vision techniques assume that given images are obtained with pin-hole camera, DFF assumes that a real aperture camera is used. Real aperture cameras have relatively short depth of field resulting in images focused only on a small distance range of a scene. For instance, Darrell and Wohn [4] use Laplacian of Gaussian pyramids in order to find best focused frame at each pixel.

There are lots of traditional Depth from focus methods [5–9]. Gaganov et al. [6] propose new SFF algorithm based on Markov Random Fields, and Mendapara et al. [7] use SUSAN operator. Mahmood et al. [8] use energy of high-frequency components in S transform and Mahmood et al. [9] propose to use 3D anisotropic nonlinear diffusion filtering for accurate SFF. Recently, depth from focus (DFF) method has been improved in various ways producing more accurate depth map and high quality all-focused image. Muhammad and Choi [10] use a neural network in order to extract 3D shape of objects based on DFF. In Reference [11], they use neural networks to learn the shape of focused image surface (FIS). In Reference [12,13], the authors use dynamic programming optimization technique to get depth (3D shape) of the scene from set of focused images. Their method works significantly faster than previous FIS algorithms. In Reference [14], the authors propose local search algorithm for SFF problem to reduce computational complexity. In Reference [15], the authors propose focus on measurement method based on an optical transfer function, which is implemented in Fourier domain for 3D shape recovery. Sun et al. [16] employ the entropy of high frequency bands as the amount of blur combining texture information for improved performance. Liu et al. [17] propose a semi global DFF approach that enforces the adaptive smooth constraint on reconstructed depth of the scene. Meoller et al. [1] propose variational approach using an efficient non-convex minimization scheme to produce depth map. Suwajanakorn et al. [18] formulate uncalibrated DFF problem and propose a new focal-stack aligning algorithm to estimate depth of given scene using hand held cameras.

In Reference [19], the authors propose a new focus measurement that is robust to noise with higher accuracy in focus measurement. They present Ring Difference filter (RDF) by inserting a gap and looking at the pixels that are located farther away from the point of interest (POI). They extended their work [20] by proposing RDF-based cost aggregation that utilizes both local and non-local characteristics by inheriting the structure of RDF. Recently, Zhiqiang et al. [21] proposed depth recovery framework including depth reconstruction and refinement process. They use non-local matting Laplacian prior and variance based confidence level computation. It is able to produce depth map robust to texture-less regions and give more clean edges. Hazirbas et al. [22] propose auto-encoder-style network to predict depth from focal stack. In order to train their network, they create 12-Scene Benchmark dataset. For the encoder part, they use VGG-16, a popular deep neural network for object recognition [23], without fully connected layers, and for decoder network they simply use flipped structure of the encoder network. It provides sharpness map (feature map) for each frame separately with single output depth map. However, such deep learning approach has limitations, such as fixed number of input frames (it depends on training samples).

However, traditional DFF methods suffers from unstable approximation results with texture-less regions and object boundary regions. Indeed, texture-less region shows limited clue for the estimation of focus amount from only its visual appearance and object boundary contains sudden changes in depth that making patch based focus estimation difficult. In order to resolve such limitations, we suggest maximizing depth inference from neighbor pixels and neighbor frames globally optimizing overall changes of focus amount along the spatial and frame domains. In this work, we propose a new DFF method using single color camera. We investigate the appearance changes in spatial and frequency domains over differently focused image frames. In spatial domain, visual appearance change of low-textured region along the frames is difficult to be observed. In frequency domain, however, low-textured region gives slightly better observation over frames. Using both spatial and frequency domain observations, we can get robust depth estimation at each pixel even in low-textured regions. In order to achieve sub-frame level accuracy in depth estimation at each pixel, optimal location of best focused frame is estimated by fitting a second order (quadratic) polynomial curve on the dissimilarity measurements. Based on the estimated initial depth, we build a uniformly focused image set by pixel-wise adjustment of depth level. In other words, all-in-focus image and following uniformly out-focused images are created. We perform our depth estimation iterative manner on this uniformly focused image set that refine our depth estimation. During depth estimation process we create confidence map that helps to fill erroneous depth points caused by textureless uniform regions. After filling erroneous points, we use guided filter [24] using predicted depth to build its all-in-focus image with which we finally create cleaner depth image. Each estimated depth point is considered as 3D point, and surface normal of the point is computed from its 8 neighbors [25]. We interpolate them using Algebraic Point Set Surface(APSS) [26], and, finally, using marching cubes algorithm [27], we create polygonal surface representation.

Our main contributions are the proposal of a new focus measure that uses both spatial and frequency domain information, creating a uniformly focused image set (UFIS) to achieve more accurate depth, as well as hole-filling process that fixes erroneous depth values caused by textureless region. Figure 1 illustrates the pipeline of our proposed method. We tested our proposed method with Light Field dataset (LFSD) [28], dataset from smartphone, synthetic, and teeth datasets (medical). We compare our method with existing focus-measure [29–37], and state-of-the-art DFF [1,18–20,22,38,39] methods quantitatively (Table 1) and qualitatively.

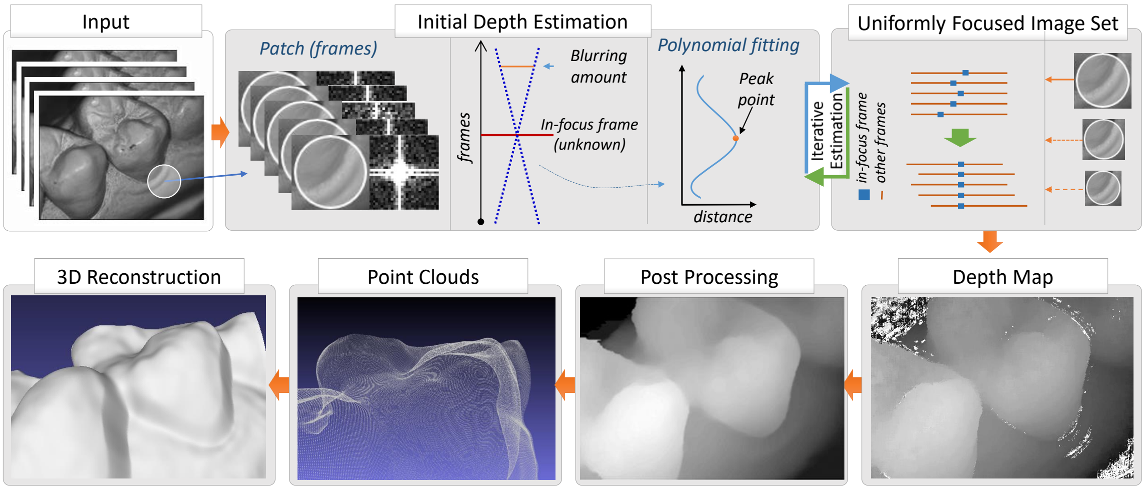


Figure 1. Pipeline of our proposed depth from focus method.

Table 1. Focus measure operators and depth from focus (DFF) methods with abbreviation and year of publication.

Focus Measure Operators	Abbr.	Year	Depth from Focus	Abbr.	Year
Gaussian derivative [29]	**GDER**	2000	DFF from mobile phone [18]	**DFM**	2015
Variance of laplacian [30]	**LAPV**	2000	Variational DFF [1]	**VAR**	2015
Image contrast [31]	**CONT**	2001	Noise robust RDF [19]	**RDF**	2017
Wavelet sum [32]	**WAVS**	2003	Composite Focus Measure [38]	**CFM**	2017
Variance of Wavelet [32]	**WAVV**	2003	PAD method of multipliers [39]	**PAD**	2018
DCT Energy measure [33]	**DCTE**	2006	Deep Depth from Focus [22]	**DDFFNet**	2018
Wavelet ratio [34]	**WAVR**	2006	Fast and noise robust RDF [20]	**F-RDF**	2019
DCT Energy ratio [35]	**DCTR**	2009			
Diagonal Laplacian [36]	**LAPD**	2009			
Steerable filters-based [37]	**SFIL**	2009			

2. Proposed Depth from Focus

The proposed method consists of three steps. In first step initial depth of each pixel is estimated. The second step is iterative refinement building uniformly focused image set. The last step is post-processing that improves depth estimation in both textureless and object boundary regions.

2.1. Initial Depth Estimation

In order to find best-focused frame of each pixel, we calculate neighborhood patch dissimilarity. Small window size causes noisy depth due to the lack of enough observation for focus estimation. On the other hand, a bigger patch loses details in depth. In our work, we use adaptive patch size at each iterative estimation. In addition, we use a circular window to minimize the artifact of estimated depth.

A circular window creates less perceivable noise. Figure 2 shows the difference of 3D reconstruction result between square and circle window.

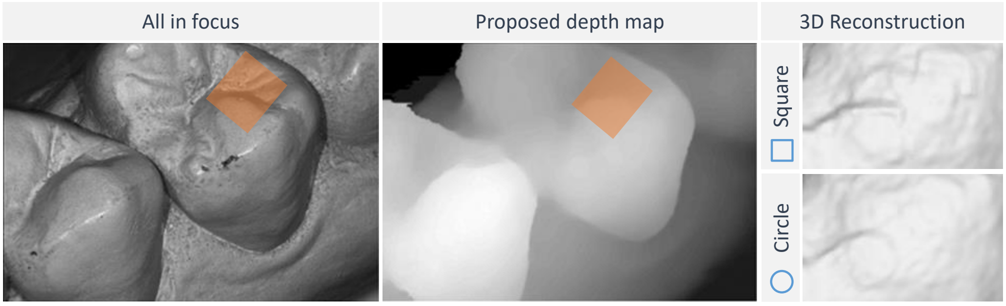


Figure 2. From left to right: all-in-focused image, depth map, and 3D reconstruction results using circle and square windows (highlighted part zoomed-in). Three-dimensional reconstruction result shows that circle window produces less artifacts.

Using patch dissimilarity to find the best-focused frame, we chose a reference frame arbitrarily among focal-stack. Then, we calculate the normalized average absolute difference of patches between the reference frame and all other frames at each pixel. The best-focused frame becomes a local maximum in the difference. The best-focused frame detection task for each pixel becomes a simple local maximum detection task. However, if the best-focused frame is identical or very close to the reference frame, the best-focused frame cannot be detected easily. To avoid such a singular case, we blur the reference frame before the difference computation. The blurring amount of the reference frame is selected not to have a similar blur amount in all other focused image frames (Figure 3a). We use Gaussian blurring assuming that it simulates optical blur well in out-focus images.

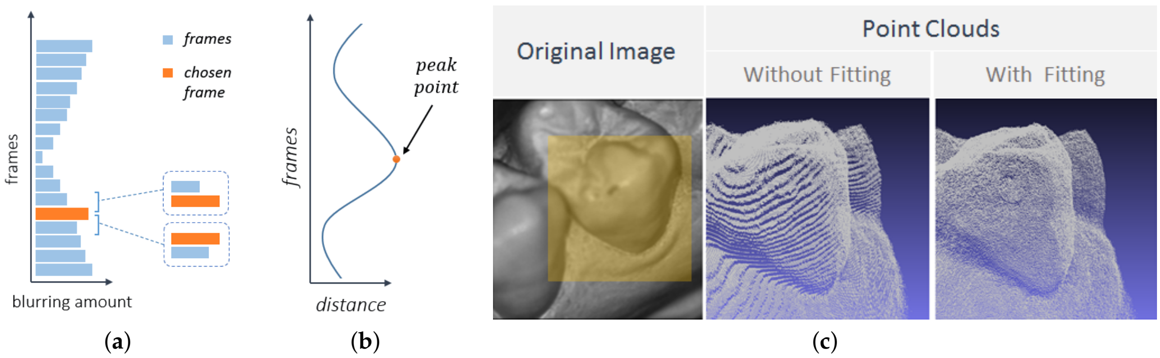


Figure 3. (**a**) Adding blur for reference frame (yellow: reference frame, blue: other frames); (**b**) difference between reference frame and other frames (curve form); (**c**) comparison of obtained 3D point clouds with and without curve fitting.

Normalized average absolute difference between the reference frame and kth frame is calculated on both spatial and frequency domain information of the given patch using Equation (1) for spatial domain and Equation (2) for frequency domain:

$$p_k^{x,y} = \sum_{i=x-w}^{x+w} \sum_{j=i-w}^{y+w} \left(\left| R_{i,j} - F_{i,j}^k \right| * mask_{i,j} \right), \quad (1)$$

$$\bar{p}_k^{x,y} = \sum_{i=x-w}^{x+w} \sum_{j=i-w}^{y+w} \left(\left| \bar{R}_{i,j} - \bar{F}_{i,j}^k \right| \right), \quad (2)$$

where x and y are pixel indexes, w is window size, R is reference frame in spatial domain, $\bar{R}$ is reference frame (neglect edge effects using mask) in frequency domain, k is index of frames ($1 \leqslant k \leqslant n$, n is number of frames in focus-stack), F is kth frame in spatial domain, $\bar{F}$ is kth frame in frequency domain, and $mask$ is binary circle mask (to use circle window). Then, we create distance vector for each index of frame with applying weight and normalization for Equations (1) and (2) using Equation (3):

$$d_k = \alpha \left(p'_k\right) + (\alpha - 1) \left(\bar{p'}_{k'}\right) \tag{3}$$

where the $apostrophe$ mark means that variable is normalized between 0 and 1, and α is a weight with which we can assign a preference on either spatial domain clue or frequency domain clue (Figure 4).

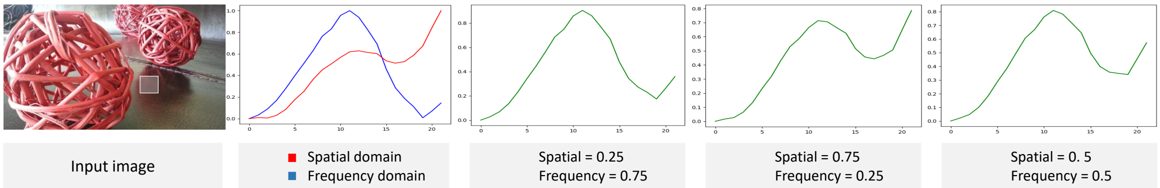
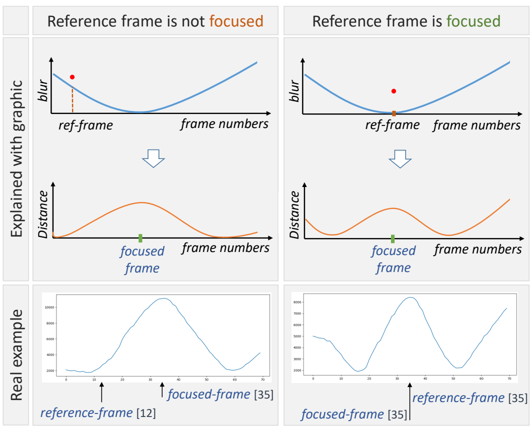
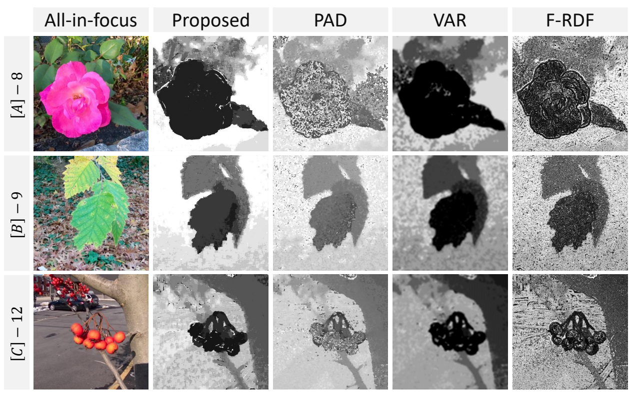


Figure 4. Difference between reference frame and other frames on less textured part of the image (marked with rectangle). Each graph: x axes—frame index; y—axes difference between reference frame and others (normalized between 0 and 1). From left to right: one of the input image, frame difference on frequency (blue) and spatial (red) domain, third to fifth: summation results after applying weight to both domain.

Final step is finding the index of maximum distance ($max(d_k)$) which indicates potential best-focused frame. When we calculate the difference between reference-frame and other frames, the best-focused frame may gives bigger distance than its neighbors (Figure 5a). Note that, in DFF, the best-focused frame's index indicates the depth of the corresponding point. In order to be more accurate by estimating sub-frame level best focused frame, we use parameterized quadratic curve fitting [40] around the maximum index (Figure 1 (curve fitting) and Figure 3b). We use second order polynomial $f(x) = a_0 + a_1x + a_2x^2$ to fit on d_k. Figure 3c shows our depth estimation with and without curve fitting.



(**a**) Choosing reference-frame arbitrary (real and synthetic graphic). In real example, it showed distance between reference-frame and other frames. *Red — point*—blur amount of reference-frame (likely), *green — mark*—focused-frame



(**b**) Initial depth map of proposed and recent methods. Initial depth map shows that the proposed focus measure method is more robust than state-of-the-art methods, such as PAD [39], VAR [1], and F-RDF [20]

Figure 5. (**a**) Choosing reference frame arbitrary; (**b**) initial depth result of DFF methods.

2.2. Iterative Depth Refinement

Initial depth is estimated by investigating the appearance changes in both spatial and frequency domains in the previous section. To obtain more accurate result, we build a *uniformly focused image set* (UFIS) by taking pixels from the focus stack using resulting depth values. In other words, an all-in-focus image and following uniformly out-focused images were created. Then, using UFIS, we improve our depth estimation iterative manner. Let n be a number of frames of UFIS. n is an odd number, and the central frame c should be an all-in-focus frame generated by the initial depth result (if n = 7 then c = 4, if n = 5 then c = 3). We fill all other frames of UFIS using the following equation:

$$T^{i}_{x,y} = F^{f}_{x,y}, \tag{4}$$

where x and y are pixel indexes, T is UFIS (synthetic focal stack), i is an index of the reconstructed frame number of T, F is initial focus stack, and f is the corresponding frame number calculated by Equation (5) using initial depth estimation result:

$$f = round(D_{x,y}) + (i - c), \qquad f = \begin{cases} d & 0 \leq d \leq n \\ n & d > n \\ 0 & \text{otherwise} \end{cases}, \tag{5}$$

where i is an index of the reconstructed frame (same with Equation (4)), $D_{x,y}$ is estimated depth at $[x, y]$ position (same position with Equation (4)), and c is the central frame index of T reconstructed frame.

After UFIS is constructed, we repeat the initial depth estimation process with decreasing window size. Based on the result of *Iterative Depth Refinement*, we update the depth result. We continue this iteration process with a smaller window size until the changes in the depth value at each iteration falls below a stopping criteria. We update depth result at each iteration using Equation (6):

$$D_{x,y} = D^{p}_{x,y} + \left(d_{x,y} - c\right), \tag{6}$$

where $D^{p}_{x,y}$ is previous depth at (x,y) location, $d_{x,y}$ is depth result in the current iteration, and c is center frame index.

Figure 6 shows the results of 4 iterations. The first image is one of input frames, second is a point cloud of the scene from initial depth estimation, and A,B,C,D images show the point cloud of each iteration: green color is first iteration with window size 7×7, orange color is second iteration with window size 5×5, blue color is third iteration with window size 3×3, and purple color is fourth iteration with window size 1×1. Image A shows the point cloud results of all iterations, and the green point cloud shows more flat results at the highlighted region (in Figure 6). Image B shows second, third, and fourth iterations' point cloud result, and the yellow point cloud has provided a more flat result than blue and purple (in Figure 6). Image C shows that the blue point cloud has more flat regions than purple (in Figure 6). It shows that each iteration point cloud becomes more accurate (in highlighted more curved case). UFIS process tries to reach an optimal result with corresponding initial (or previous) depth estimation.

Figure 7 describes the creation of our UFIS.

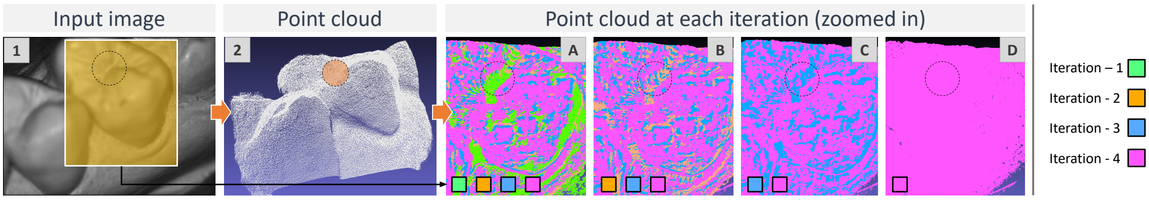


Figure 6. Input image, point cloud(without post-processing) and zoomed-in version of point cloud result of each iterations. Iteration 1-green, 2-yellow, 3-blue, 4-purple. A—includes all iterations' result, B—includes 3 iterations' result, C—includes 2 iterations' result, D—includes last iterations' result. First iteration's (green) point cloud cannot provide detailed depth, next iterations' result shows more precise depth. Rectangle color of each point cloud result means that iterations are visible in that figure. Full view of depth map and point cloud shown in Figure 1.

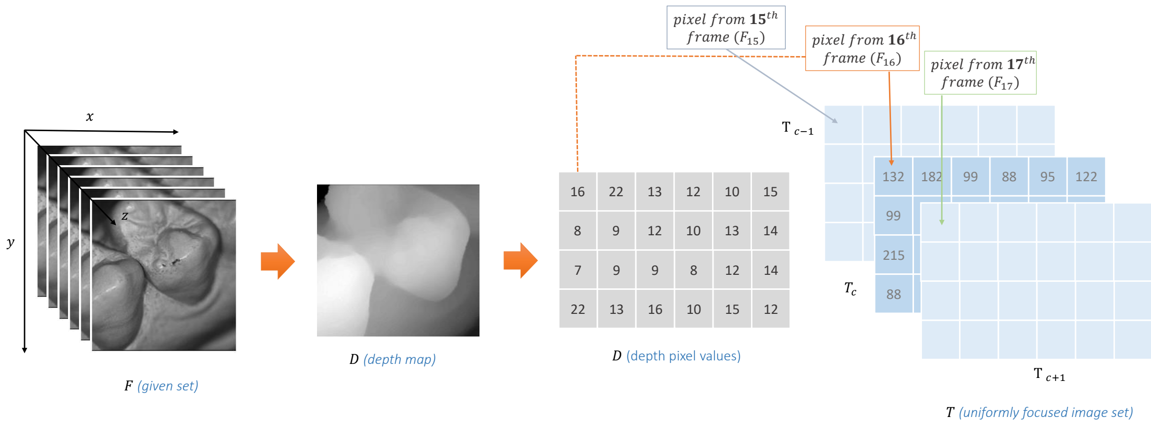


Figure 7. Iterative depth refinement creating uniformly focused image set.

2.3. Textureless Region and Post Processing

Estimating depth on the textureless region is one of the challenging problems in DFF. Most existing DFF methods fail in textureless regions because of the lack of enough information to observe. Figure 8 describes two different parts of the input scene: textured and textureless regions.

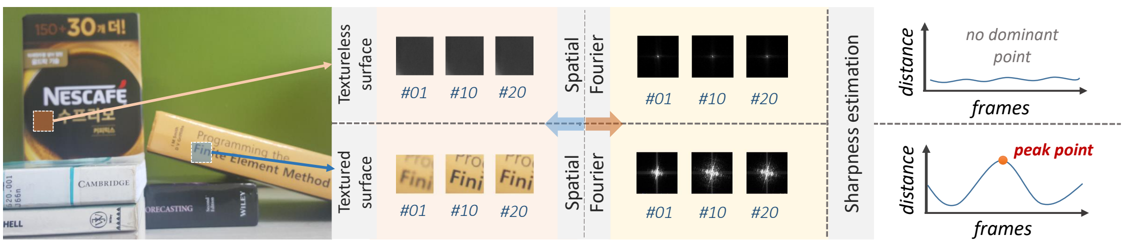


Figure 8. Choosing peak-point(best-focused frame) of normalized absolute difference between the reference frame and all other frames from textured and textureless parts of the scene.

The image shows that textured surface has more information to observe, and the textureless surface has a lack of enough information to chose the best-focused frame. In Figure 9, the depth estimation result has a bunch of erroneous points on textureless regions. To fix these erroneous points, we have to use depth information from its neighbors.

In order to identify textureless region, we investigate all-in-focused image (generated using depth map) and find inaccurate (noisy) regions from depth map, as well. We create two binary masks:

- Textureless Region Mask (TRM): textureless regions of all-in-focused image, and
- Inaccurate Depth Mask (IDM): erroneous depth regions on depth map.

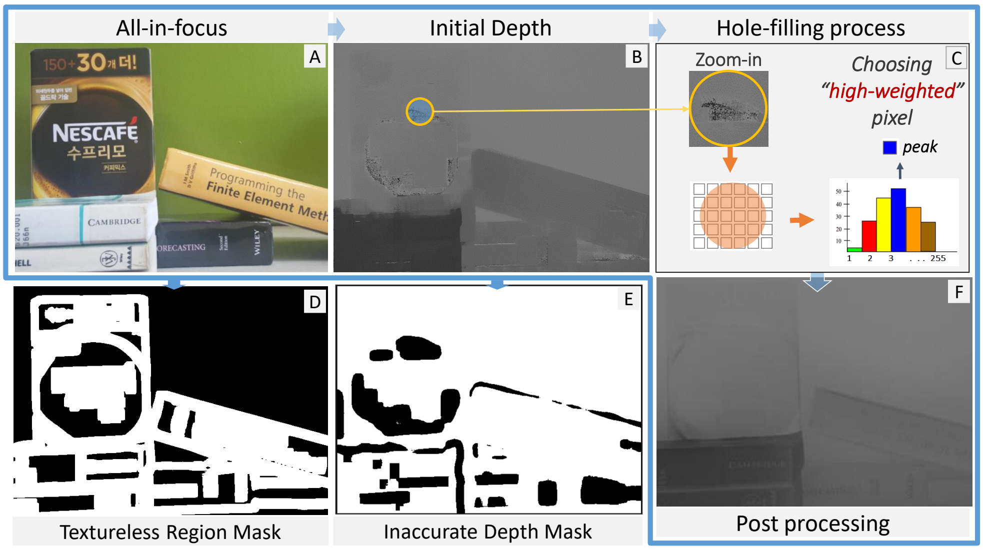
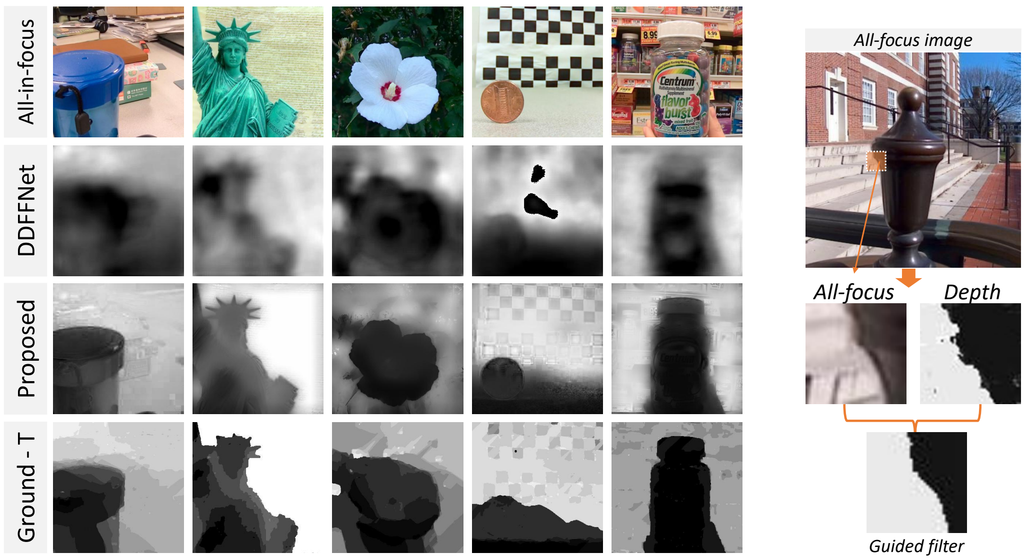


Figure 9. A hole-filling process on textureless regions that produce erroneous depth values (note: coffee box and the wall (background) are close to each other, so they have almost similar depth). (**A**,**B**) images are all-focused-image and initial depth map; (**C**): Hole-filling process; (**D**): Textureless Region Mask; (**E**): Inaccurate Depth Mask; (**F**): Post processing.

TRM mask. Textureless regions have lower variance in color. In order to create *TRM* mask, we scan a window over all-in-focused image calculating color variance within the window. If the variance is lower than a threshold, we mark the point as textureless. Threshold is the squared value of the window size ($th = w^2$, w is window size). Threshold value is chosen empirically as the value that produces a better result than other variations when pixel values are in between 0 and 255. Figure 9D describes *TRM* mask (black-color: textureless region).

IDM mask. Inaccurate depth points which are from textureless part have higher variance while accurate depth points have lower variance. Based on this assumption we create *IDM* mask. Generating *IDM* mask process is similar to *TRM*. If the variance is higher than a threshold, we mark that point as an inaccurate depth point. Figure 9E describes *IDM* mask (black-color: inaccurate depth). *IDM* sometimes finds inaccurate regions even at the correct depth around edges of objects. We overlap both masks to figure out textureless regions. We indicate *erroneous depth from textureless* regions as (*EDfT*). Finally, hole-filling is performed with *EDfT* regions. We replace *EDfT* points by its neighbors of high-score-confidence. Depth confidence gets higher if the peak point in the sharpness estimation is more greater than its neighbors in the focus measure process. Based on this idea, we created a score-confidence map in the depth estimation process.

After fixing depth values of textureless region, we perform post-processing for getting cleaner depth around the object boundary using a filter called guided filter [24]. The guided filter investigates intensity changes in an all-in-focus image (Figure 10b). Finally, we scale estimated depth values by scaling factor s to convert frame index to physically meaningful depth value. Median and Gaussian filters are applied, in turn, to eliminate any remaining noisy value. After obtaining depth information of the scene, we can get point clouds of the scene, as shown in Figure 1 (point clouds).



(**a**) (**b**)

Figure 10. (**a**) Comparison with Deep Depth from Focus (DDFF)Net on LFSD dataset [28]. All samples have 10 frames. First-row: input image, second: results from DDFFNet [22], third: our result, last-row: ground-truth. (**b**) Effect of guided filter [24] on the depth map: it helps keeping edges.

3. Experimental Results

3.1. Experimental Setup

In experimental evaluation, we use 11×11 window size and make it circular mask with corresponding binary mask. In the iteration step, window size is gradually reduced until it reaches 1×1 size. We use $alpha = 0.5$ on Equation (3). When we choose peak-point to find focused frame, we use second order polynomial fitting. Reference frame number is $round(n/3)$ (n is number of frames in stack), and then we apply Gaussian-blurring (size: 11×11).

3.2. Error Metrics

In order to evaluate quantitatively, we use following several measurements, which are used in previous works [41,41–48]: Root mean squared error (linear) (Equation (7)), Root mean squared error (log) (Equation (8)), Absolute relative error (Equation (9)), Squared relative error (Equation (10)), and Accuracy (% of y such that, Equation (11)):

$$RMS = \sqrt{\frac{1}{T} \sum_{i,j} \left| y_{i,j} - y^*_{i,j} \right|^2}, \tag{7}$$

$$\mathbf{RMS}(log) : \log_{10} = \sqrt{\frac{1}{T} \sum_{i,j} \left| \log y_{i,j} - \log y^*_{i,j} \right|^2}, \tag{8}$$

$$rel = \frac{1}{T} \sum_{i,j} \left| y_{i,j} - y^*_{i,j} \right| / y^*_{i,j}, \tag{9}$$

$$srel = \frac{1}{T} \sum_{i,j} \left| y_{i,j} - y^*_{i,j} \right|^2 / y^*_{i,j}, \tag{10}$$

$$\left(\frac{y_i}{y^*_i}, \frac{y^*_i}{y_i} \right) = \sigma < th, \tag{11}$$

where y is depth map, y^* is ground-truth, and T is depth-pixels.

3.3. Qualitative and Quantitative Evaluation

In order to evaluate our work, we use four different datasets: LFSD dataset [28], Lytro dataset [49], dataset from mobile [18], and our synthetic and medical (teeth) datasets. We compare our method with existing focus-measure focus-measure [29–37] and state-of-the-art DFF [1,18–20,22,38,39] methods.

Figure 11 shows results on the LFSD dataset [28]. The LFSD dataset [28] provides focal stacks captured by Lytro Illum camera with corresponding depth map. It has 5 to 12 focus image frames for each subject. The depth map has darker intensity for farther depth. The proposed method estimates depth of each sample reasonably well.

Table 2 shows quantitative evaluation of our method and previous focus measure methods (to get depth result we used our post-processing without filling textureless region) and state-of-the-art DFF methods.

Table 2. Quantitative evaluation on the LFSD dataset [28].

Methods		Lower Is Better				Accuracy (Higher Is Better)		
		RMS	**Log RMS**	**Abs rel**	**Sqr. rel**	$\delta = 1.25$	δ^2	δ^3
Focus measure operators	**GDER** [29]	9.964	0.519	1.279	0.997	49.687	69.968	78.326
	LAPV [30]	9.608	0.417	1.553	0.922	56.028	76.290	82.680
	CONT [31]	9.890	0.503	1.523	0.988	52.363	71.991	80.187
	WAVS [32]	9.665	0.404	1.607	0.926	56.013	76.930	83.064
	WAVV [32]	9.668	0.391	1.575	0.921	56.153	77.204	83.848
	DCTE [33]	9.817	0.458	1.447	0.962	52.603	73.081	81.620
	WAVR [34]	9.684	0.401	1.583	0.928	56.090	76.807	83.336
	DCTR [35]	10.091	0.578	1.194	1.056	39.648	61.689	72.927
	LAPD [36]	9.706	0.419	1.585	0.944	55.878	76.445	82.596
	SFIL [37]	9.973	0.558	1.281	1.002	47.758	67.962	76.599
DFF	**VAR** [1]	10.110	0.413	1.553	1.050	51.442	59.558	80.715
	RDF [19]	9.865	0.443	1.433	0.971	52.623	72.891	82.676
	PAD [39]	10.227	0.506	1.178	1.113	43.446	62.380	74.864
	F-RDF [20]	9.850	0.444	1.434	0.961	52.555	72.559	82.693
	Proposed	8.915	0.338	1.223	0.843	63.199	80.653	86.495

In order to evaluate our focus measurement without post-processing, we collected synthesized focus image set (focus level changed by adjusting camera focus) from graphics models, including ground truth depth. Figure 12 shows sample depth estimation results compared with ground truth. Depth images are normalized with min/max value corresponding to ground-truth, and depth images are from initial depth result applying median filter (size: 3×3) without using post-processing. In this test, we vary patch size (Table 3). Each synthetic subject consists of 31 focus image frames with size 405×720. Accuracy in Table 3 is measured by average absolute difference of depth values. Each window size shows RMSE (Root Mean Squared Error) of the related sample, and the last row show average absolute difference of them. 11×11 window size shows best estimation accuracy, however it depends on the complexity of the shape of test objects. If objects have complicated surface shape, smaller window size will work better, but it may cause a noisy result (as shown 5×5 window size's result).

Figure 11. Experimental results on the LFSD dataset [28]. (**A**)—bench (6-frames), (**B**)—flower (10), (**C**)—lollipop-toy (7), (**D**)—flower-2 (9), (**E**)—flower-3 (12), (**F**)—marble-stone (12), (**G**)—door-handle (12), and (**H**)—liberty statue (10). The first row shows all in focus images, 2–9 rows are results from previous focus measuring and state-of-the-art DFF methods, 10-rows show reconstructed depth maps of our proposed method, and last row shows show given depth from Reference [28].

Table 3. Average absolute difference of estimated depth.

Sample	Window Size				
	5 × 5	**11 × 11**	**15 × 15**	**21 × 21**	**31 × 31**
Books	23.88	22.79	23.70	23.51	23.71
Balls	9.08	7.43	9.13	9.44	8.97
Plants	12.09	14.08	15.41	19.96	22.08
Average	15.01	14.76	16.08	17.63	18.49

We constructed 3D mesh model of target object using our teeth dataset (Figure 13). These results show that our proposed method prove the potential of single color camera-based 3D reconstruction. In our teeth dataset has 100 focus image frames for each subject taken by dental intraoral camera. Each estimated depth point is considered as a 3D point, and surface normal of the 3D point is computed from its 8 neighbors [25]. After that, we interpolate them using Algebraic Point Set Surfaces (APSS) [26] method. It creates a smooth surface using local moving least-squares (MLS) approximations of the data [50]. Finally, marching cubes algorithm [27] creates polygonal surface representation (Figure 13). This evaluation result shows the potential of the proposed method in medical 3D imaging application with simple color camera.

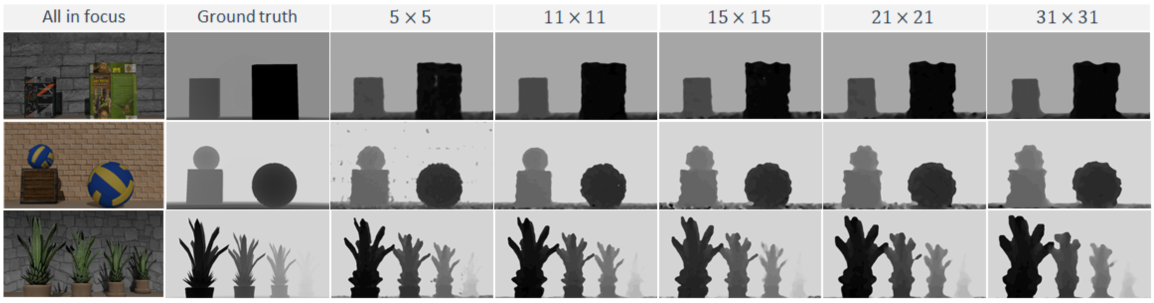


Figure 12. Depth estimation of our method on synthetic dataset.

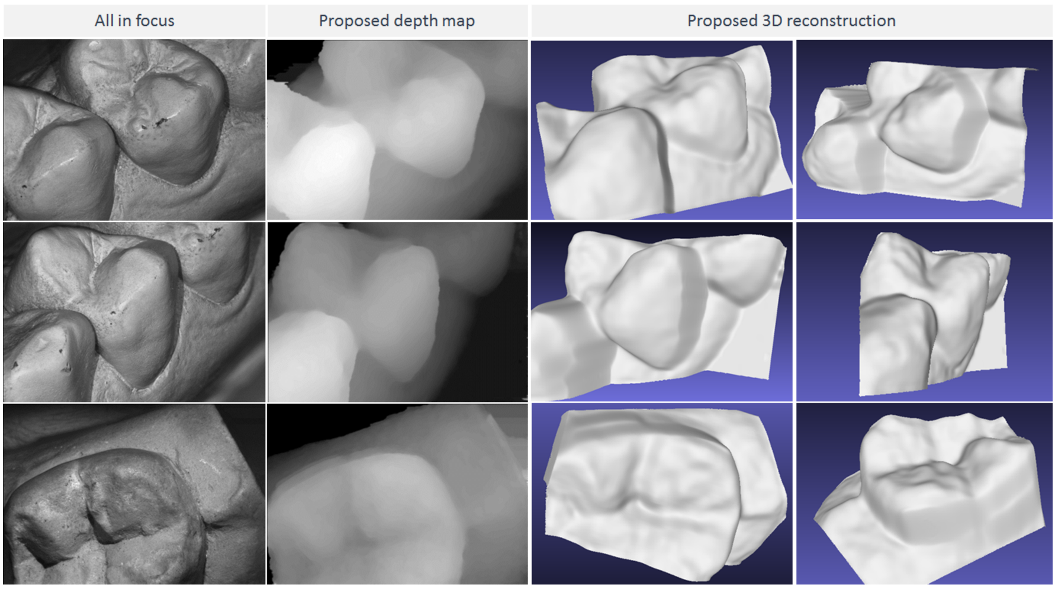


Figure 13. Our depth estimation and 3D reconstruction results on teeth dataset (extended result of Reference [51]).

We compare our model with Deep Depth from Focus (DDFF)Net [22]. Unfortunately, their model requires 10 images with fixed size. LFSD dataset [28] has several samples with 10 frames. So, we run our method with same inputs (10 frames). Figure 10a shows result comparison with DDFFNet [22]. DDFFNet provided blurry and poor depth result, which barely kept structure and depth of the scene on LFSD dataset [28]. Our proposed method provides better depth estimation, which is clean and accurate.

3.4. Comparison with the State-of-the-Art

In this section, we compare the proposed method with state-of-the-art DFF methods, such as VAR [1], DFM [18], RDF [19], Composite Focus Measure (CFM) [38], PAD [39], and F-RDF [20]. For experimental evaluation, we use the LFSD dataset [28], the Lytro dataset [49], and the dataset from mobile [18].

Figure 14 shows depth and point-cloud results of VAR [1], RDF [19], PAD [39], F-RDF [20] and proposed method on "Buddha" sample. The proposed method provide more clean and accurate depth, especially around object boundaries. The last row shows a zoom-in part of the highlighted region where the proposed method has more accurate depth than others.

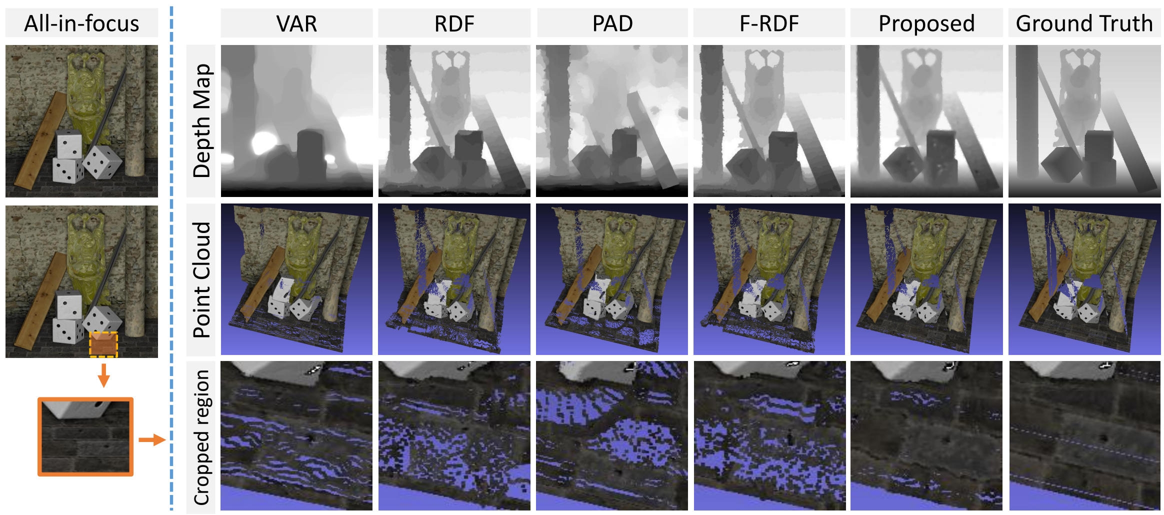


Figure 14. Comparison of point cloud of the scene with state-of-the-art methods. First row shows depth map, second: point cloud results, and last row shows cropped (zoomed-in) region of the scene. The proposed method provide with correct depth value when most methods fail (at cropped region).

Figure 15 shows peak signal-to-noise ratio (PSNR) between predicted depth and ground truth. In order to compute $PSNR$, the block first calculates the mean-squared error using Equation (12). Then, the block computes the $PSNR$ using Equation (13)

$$MSE = \frac{\sum_{M,N} [I_1(m,n) - I_2(m,n)]^2}{M * N}, \tag{12}$$

$$PSNR = 10 \log_{10} \frac{R^2}{MSE}. \tag{13}$$

In Equations (12) and (13), M and N are the number of rows and columns in the input images. R is the maximum fluctuation (the maximum possible pixel value of the image) in the input image data type. This figure shows that quality of the proposed method outperforms state-of-the-art methods.

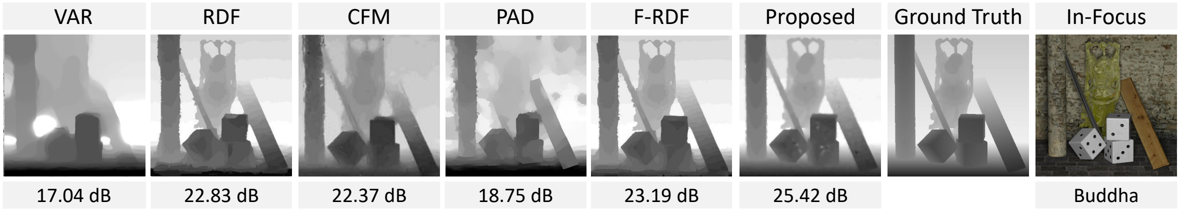


Figure 15. Quantitative evaluation on 'Buddha' synthetic sample (with 25 frames) with state-of-the-art methods. From left to right: Variational DFF (VAR) [1], Ring Difference filter (RDF) [19], Composite Focus Measure (CFM) [38], PAD method of multipliers (PAD) [39], Fast and noise robust RDF (F-RDF) [20], proposed, ground-truth, and in-focus image. The proposed method perform better than recent methods that visible in the images and reflected in the peak signal-to-noise ratio (PSNR) (in dB) shown below the each depth result. Note that depth image and PSNR value for CFM [38] was cropped from original paper.

Figure 16 shows qualitative performance of the proposed method on LFSD dataset [28] over the recent methods, such as VAR [1], RDF [19], PAD [39], and F-RDF [20]. Each sample has 5 to 12 frames of size = 1080 × 1080. Our proposed method provided clean depth with detail and clean edges, especially on samples B (branch of the tree), D (wire on the window), F (boundary of the statue), and H (pens).

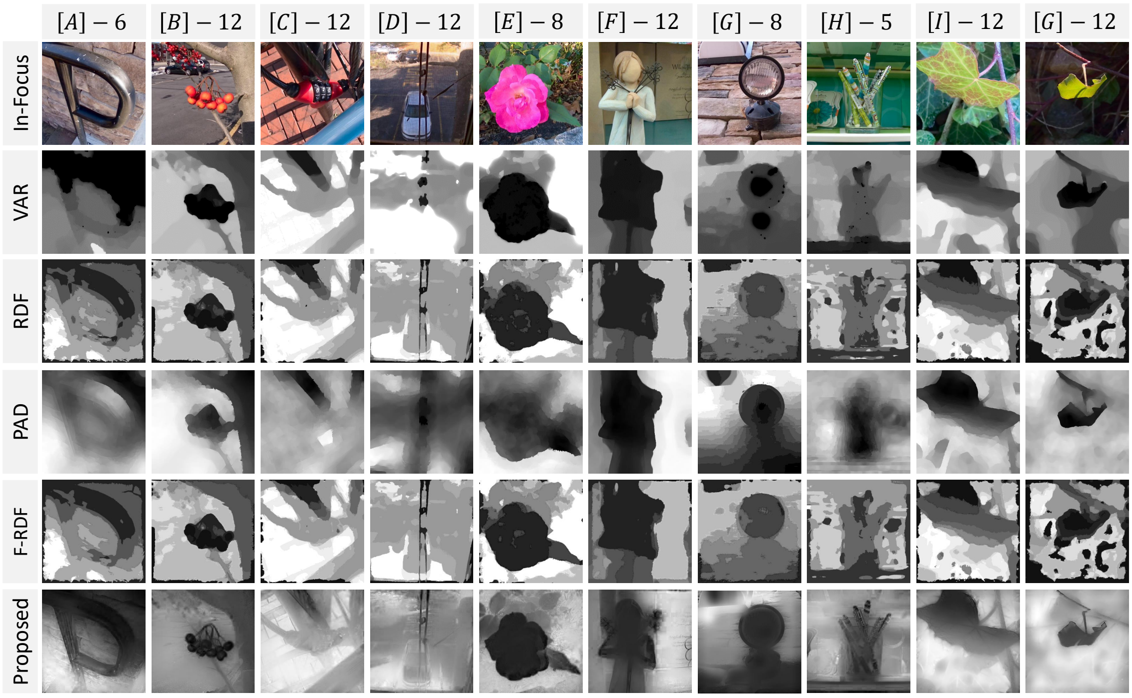


Figure 16. Qualitative evaluation on LFSD dataset [28] with state-of-the-art methods. The top of each sample has the number of used frames. The proposed method provide more accurate depth with clean edges, even at a complex scene, such as shiny object (**G**), more clean details on (**B**,**D**,**F**–**H**), and proposed method kept structure of the scene on (**A**,**C**,**E**) samples better than other methods.

Figure 17 shows results of VAR [1], RDF [19], PAD [39], F-RDF [20], and the proposed method on Lytro dataset [49]. Our proposed method gives clean depth results, especially with sample E around the boundary of the lock.

We evaluated the proposed method on many challenging samples. In Figures 18a and 19, proposed method provide clean and detailed depth even on complex region of the scene. In Figure 19, the proposed method provided cleaner edges than other methods on all samples, and second sample has leaf that only our method could able to provide accurate depth. Figure 18a has the complex part of the scene that is highlighted; "Toy" sample has small details that the proposed method provided accurate depth with clean edges, and "Buddha" sample has a column that shows proposed method provided accurate depth, while other methods failed, and the proposed method has more a accurate result on the "Flower" sample with small details.

In the evaluation on mobile dataset [18] shown in Figure 20, we use three test-samples and compare our work with state-of-the-art methods, such as DFM [18] and CFM [38]. The first row shows *"plants"* sample with 30 frames, second: *"fruits"* sample with 23 frames, and third row shows *"books"* sample with 14 frames. Our method finds more accurate depth than prior works, especially around the edges on all samples. On the *"books"* sample, DFM [18] failed to provide clean depth on background because of textureless region (background has black color and no texture), and CFM [38] provided a noisy result, while the proposed method provided clean depth even at textureless (background) regions. These results prove that the proposed method can provide a clean and accurate edge, even on textureless regions.

Experimental evaluations show that our method gets improved depth estimation results over existing methods. PAD [39] method is less robust to textureless regions and suffers from texture-copy artifacts. DDFFNet [22] method fails to predict correct depth, providing depth is very blurry, and it has a limitation with image numbers in stack. VAR [1] can keep the overall structure of the scene but it cannot provide clean edges. RDF [19] and F-RDF [20] methods have limitations; their methods cannot provide good result if the image contains a region that is never in focus by a wide margin. The proposed method generates a depth using information extracted in both spatial and frequency domains. Fourier domain works more efficiently on low-textured, and spatial domain has better performance with textured areas. Figure 5b shows that our focus measure is more robust than state-of-the-art methods. PAD [39] and F-RDF [20] provided high noisy depth (initial depth), and VAR [1] provided a less noisy depth result, while proposed focus measure provide cleaner depth result (however, it needs post-processing).

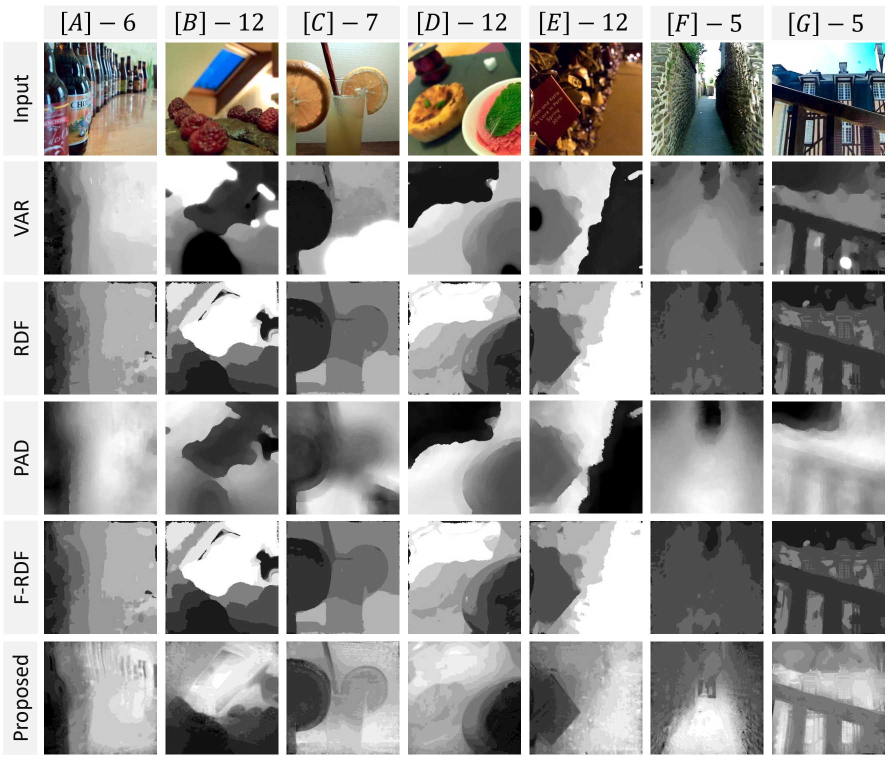
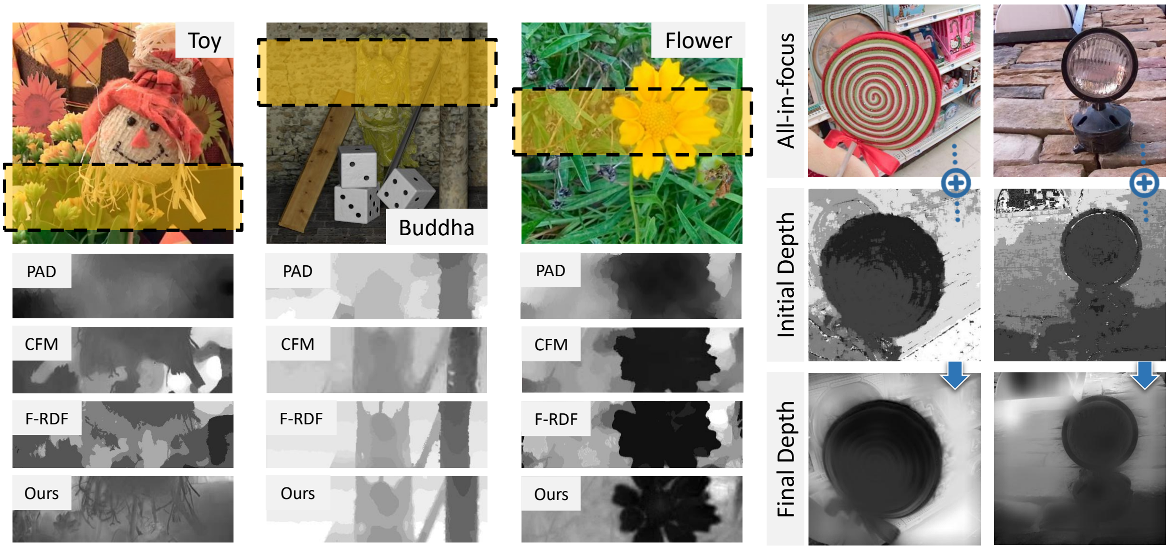


Figure 17. Qualitative evaluation on Lytro dataset [49] with recent (state-of-the-art) . The proposed method provide more accurate depth than other methods. (**A**)—beers (6 frames), (**B**)—cake (12), (**C**)—cocktails (7), (**D**)—dessert (12), (**E**)—love-locks (12), (**F**)—street (5), (**G**)—wooden-house (5).



(**a**) Qualitative evaluation on three samples (toy- 6 frames, Buddha-25, flower-12). From top to down: In-focus image (with highlighted region) and cropped part of recent methods PAD, CFM, RDF-20, and last row shows proposed method's result). Our result shows more accurate depth with details even in challenging parts of the scene. Note that depth images for CFM [38] method were cropped from the original paper.

(**b**) From top to down: in-focus frame(calculated using initial depth result), initial depth, and final depth of the proposed method.

Figure 18. (**a**) Quality of our method; (**b**) initial-depth and final-depth results.

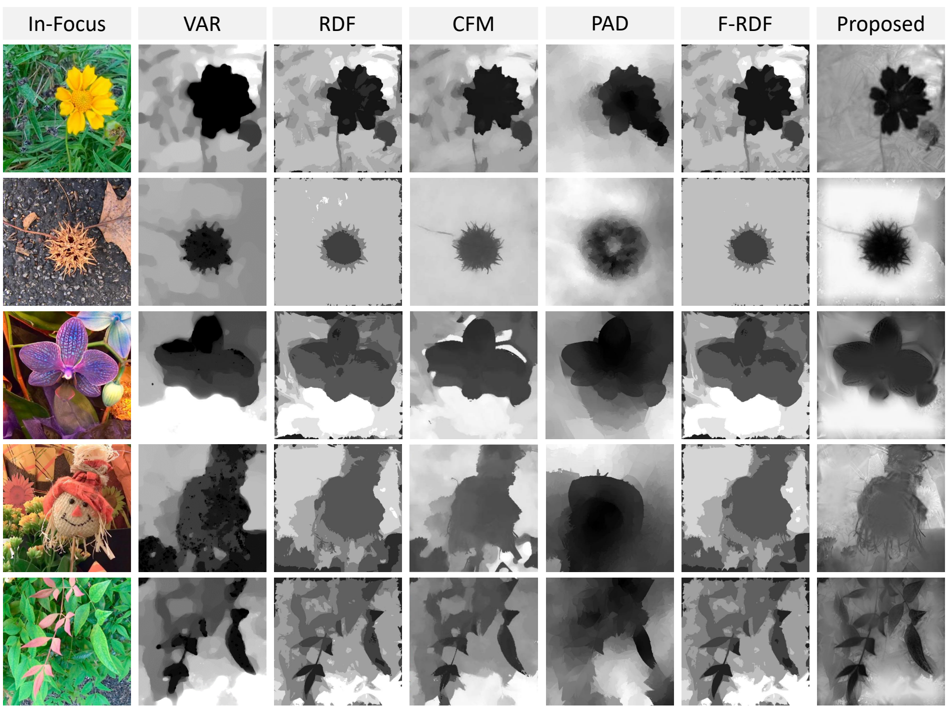


Figure 19. Qualitative evaluation on LFSD dataset [28] with recent (state-of-the-art) methods. The proposed method can provide depth with detail information even at complex scene. Note that depth image for CFM [38] method was cropped from the original paper.

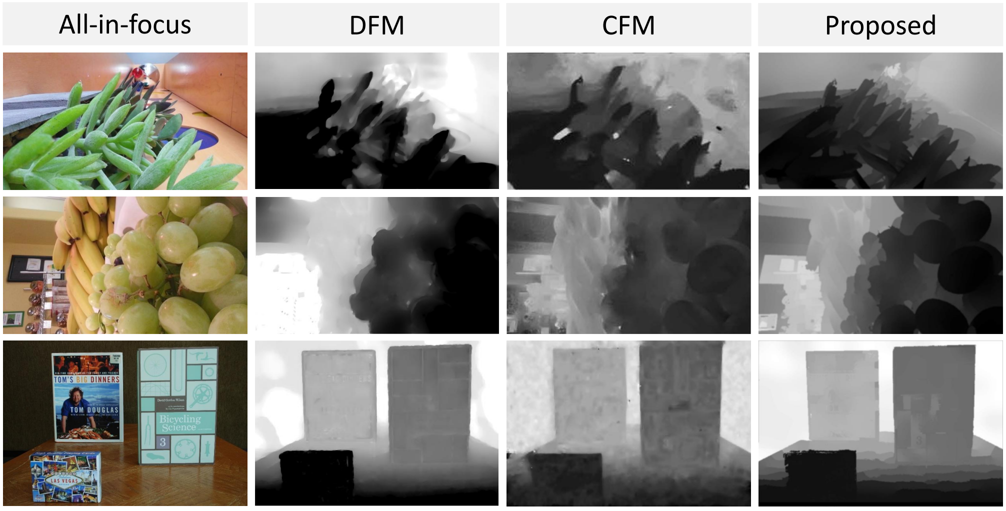


Figure 20. Qualitative evaluation (dataset from DFM [18]) with state-of-the-art methods. First column shows all-in-focus image, second: result from DFM [18], third: CFM [52], fourth: the proposed method. The proposed method provide detailed and clean edges, and third row has clean background while other methods failed on textureless region (background). Note that depth images for CFM [38] and DFM [18] methods were cropped from the original papers.

4. Conclusions

This paper proposes a new robust and accurate depth estimation method based on depth from focus (DFF) iteratively reconstructing a uniformly focused image set. We investigated the appearance changes in spatial and frequency domains in an iterative manner. We evaluated our method extensively on four public and our datasets, including a synthetic dataset. Three-dimensional modeling experiments were performed showing the potential of the proposed method in both consumer applications using a smartphone and medical applications in 3D reconstruction.

Author Contributions: Conceptualization & Methodology S.S. and S.L. Investigation, Resources, Visualisation, Data curation and Writing—original draft, S.S.; Project administration, Validation, Supervision and Writing—review & editing, S.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Moeller, M.; Benning, M.; Schönlieb, C.; Cremers, D. Variational depth from focus reconstruction. *IEEE Trans. Image Process.* **2015**, *24*, 5369–5378. [CrossRef] [PubMed]
2. Levin, A.; Fergus, R.; Durand, F.; Freeman, W.T. Image and depth from a conventional camera with a coded aperture. *ACM Trans. Graph. (TOG)* **2007**, *26*, 70. [CrossRef]
3. Zhou, C.; Lin, S.; Nayar, S.K. Coded aperture pairs for depth from defocus and defocus deblurring. *Int. J. Comput. Vis.* **2011**, *93*, 53–72. [CrossRef]
4. Darrell, T.; Wohn, K. Pyramid based depth from focus. In Proceedings of the CVPR'88, Computer Society Conference on Computer Vision and Pattern Recognition, Ann Arbor, MI, USA, 5–9 June 1988; pp. 504–509.
5. An, Y.; Kang, G.; Kim, I.J.; Chung, H.S.; Park, J. Shape from focus through Laplacian using 3D window. In Proceedings of the FGCN'08, Second International Conference on Future Generation Communication and Networking, Jeju-Island, Korea, 6–8 December 2008; Volume 2, pp. 46–50.
6. Gaganov, V.; Ignatenko, A. Robust shape from focus via Markov random fields. In Proceedings of the Graphicon Conference, Moscow, Russia, 5–9 October 2009; pp. 74–80.
7. Mendapara, P.; Minhas, R.; Wu, Q.J. Depth map estimation using exponentially decaying focus measure based on SUSAN operator. In Proceedings of the SMC 2009 IEEE International Conference on Systems, Man and Cybernetics, San Antonio, TX, USA, 11–14 October 2009; pp. 3705–3708.

8. Mahmood, M.T.; Choi, T.S. Focus measure based on the energy of high-frequency components in the S transform. *Opt. Lett.* **2010**, *35*, 1272–1274. [CrossRef]
9. Mahmood, M.T.; Choi, T.S. Nonlinear approach for enhancement of image focus volume in shape from focus. *IEEE Trans. Image Process.* **2012**, *21*, 2866–2873. [CrossRef]
10. Muhammad, A.; Choi, T.S. Learning shape from focus using multilayer neural networks. In Proceedings of the SPIE's International Symposium on Optical Science, Engineering, and Instrumentation, Denver, CO, USA, 18–23 July 1999; International Society for Optics and Photonics, Bellingham, WA, USA, 1999; pp. 366–375.
11. Asif, M.; Choi, T.S. Shape from focus using multilayer feedforward neural networks. *IEEE Trans. Image Process.* **2001**, *10*, 1670–1675. [CrossRef]
12. Ahmad, M.B.; Choi, T.S. A heuristic approach for finding best focused shape. *IEEE Trans. Circuits Syst. Video Technol.* **2005**, *15*, 566–574. [CrossRef]
13. Ahmad, M.B.; Choi, T.S. Shape from focus using optimization technique. In Proceedings of the ICASSP 2006 Proceedings, 2006 IEEE International Conference on Acoustics, Speech and Signal Processing, Toulouse, France, 14–19 May 2006; Volume 2, p. II.
14. Shim, S.O.; Choi, T.S. A novel iterative shape from focus algorithm based on combinatorial optimization. *Pattern Recognit.* **2010**, *43*, 3338–3347. [CrossRef]
15. Malik, A.S.; Choi, T.S. A novel algorithm for estimation of depth map using image focus for 3D shape recovery in the presence of noise. *Pattern Recognit.* **2008**, *41*, 2200–2225. [CrossRef]
16. Sun, H.; Zhao, Z.; Jin, X.; Niu, L.; Zhang, L. Depth from defocus and blur for single image. In Proceedings of the Visual Communications and Image Processing (VCIP), Kuching, Malaysia, 17–20 November 2013; pp. 1–5.
17. Liu, W.; Key, X.W. Semi-global depth from focus. In Proceedings of the 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR), Kuala Lumpur, Malaysia, 3–6 November 2015; pp. 624–629.
18. Suwajanakorn, S.; Hernandez, C.; Seitz, S.M. Depth from focus with your mobile phone. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–15 June 2015; pp. 3497–3506.
19. Surh, J.; Jeon, H.G.; Park, Y.; Im, S.; Ha, H.; So Kweon, I. Noise robust depth from focus using a ring difference filter. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6328–6337.
20. Jeon, H.G.; Surh, J.; Im, S.; Kweon, I.S. Ring Difference Filter for Fast and Noise Robust Depth From Focus. *IEEE Trans. Image Process.* **2019**, *29*, 1045–1060. [CrossRef]
21. Ma, Z.; Kim, D.; Shin, Y.G. Shape-from-focus reconstruction using nonlocal matting Laplacian prior followed by MRF-based refinement. *Pattern Recognit.* **2020**, *103*, 107302. [CrossRef]
22. Hazirbas, C.; Soyer, S.G.; Staab, M.C.; Leal-Taixé, L.; Cremers, D. Deep depth from focus. In *Asian Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 525–541.
23. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
24. He, K.; Sun, J.; Tang, X. Guided image filtering. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *35*, 1397–1409. [CrossRef] [PubMed]
25. Mitra, N.J.; Nguyen, A. Estimating surface normals in noisy point cloud data. In Proceedings of the Nineteenth Annual Symposium on Computational Geometry, San Diego, CA, USA, 8–10 June 2003; ACM: New York, NY, USA, 2003; pp. 322–328.
26. Guennebaud, G.; Gross, M. Algebraic point set surfaces. In *ACM Transactions on Graphics (TOG)*; ACM: New York, NY, USA, 2007; Volume 26, p. 23.
27. Lorensen, W.E.; Cline, H.E. Marching cubes: A high resolution 3D surface construction algorithm. In *ACM Siggraph Computer Graphics*; ACM: New York, NY, USA, 1987; Volume 21, pp. 163–169.
28. Li, N.; Ye, J.; Ji, Y.; Ling, H.; Yu, J. Saliency detection on light field. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 2806–2813.
29. Geusebroek, J.M.; Cornelissen, F.; Smeulders, A.W.; Geerts, H. Robust autofocusing in microscopy. *Cytom. J. Int. Soc. Anal. Cytol.* **2000**, *39*, 1–9. [CrossRef]
30. Pech-Pacheco, J.L.; Cristóbal, G.; Chamorro-Martinez, J.; Fernández-Valdivia, J. Diatom autofocusing in brightfield microscopy: A comparative study. In Proceedings of the 15th International Conference on Pattern Recognition, ICPR-2000, Barcelona, Spain, 3–7 September 2000; Volume 3, pp. 314–317.

31. Nanda, H.; Cutler, R. Practical calibrations for a real-time digital omnidirectional camera. *CVPR Tech. Sketch* **2001**, *20*, 2.
32. Yang, G.; Nelson, B.J. Wavelet-based autofocusing and unsupervised segmentation of microscopic images. In Proceedings of the 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2003) (Cat. No. 03CH37453), Las Vegas, NV, USA, 27–31 October 2003; Volume 3, pp. 2143–2148.
33. Shen, C.H.; Chen, H.H. Robust focus measure for low-contrast images. In Proceedings of the 2006 Digest of Technical Papers International Conference on Consumer Electronics, Las Vegas, NV, USA, 7–11 January 2006; pp. 69–70.
34. Xie, H.; Rong, W.; Sun, L. Wavelet-based focus measure and 3-d surface reconstruction method for microscopy images. In Proceedings of the 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems, Beijing, China, 9–15 October 2006; pp. 229–234.
35. Lee, S.Y.; Yoo, J.T.; Kumar, Y.; Kim, S.W. Reduced energy-ratio measure for robust autofocusing in digital camera. *IEEE Signal Process. Lett.* **2009**, *16*, 133–136. [CrossRef]
36. Thelen, A.; Frey, S.; Hirsch, S.; Hering, P. Improvements in shape-from-focus for holographic reconstructions with regard to focus operators, neighborhood-size, and height value interpolation. *IEEE Trans. Image Process.* **2009**, *18*, 151–157. [CrossRef] [PubMed]
37. Minhas, R.; Mohammed, A.A.; Wu, Q.J.; Sid-Ahmed, M.A. 3D shape from focus and depth map computation using steerable filters. In Proceedings of the International Conference Image Analysis and Recognition, Waterloo, ON, Canada, 27–29 August 2009; Springer: Berlin/Heidelberg, Germany, 2019; pp. 573–583.
38. Sakurikar, P.; Narayanan, P. Composite focus measure for high quality depth maps. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 1614–1622.
39. Javidnia, H.; Corcoran, P. Application of preconditioned alternating direction method of multipliers in depth from focal stack. *J. Electron. Imaging* **2018**, *27*, 023019. [CrossRef]
40. Chapra, S.C. *Applied Numerical Methods with MATLAB® for Engineers and Scientists*; Raghothaman Srinivasan; McGraw-Hill: New York, NY, USA, 2012; Chapter Curve Fitting.
41. Eigen, D.; Puhrsch, C.; Fergus, R. Depth map prediction from a single image using a multi-scale deep network. In *Advances in Neural Information Processing Systems*; Mit Press: Cambridge, MA, USA, 2014; pp. 2366–2374.
42. Liu, F.; Shen, C.; Lin, G.; Reid, I. Learning depth from single monocular images using deep convolutional neural fields. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *38*, 2024–2039. [CrossRef]
43. Eigen, D.; Fergus, R. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 2650–2658.
44. Li, B.; Shen, C.; Dai, Y.; Van Den Hengel, A.; He, M. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1119–1127.
45. Garg, R.; BG, V.K.; Carneiro, G.; Reid, I. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 740–756.
46. Honauer, K.; Johannsen, O.; Kondermann, D.; Goldluecke, B. A dataset and evaluation methodology for depth estimation on 4d light fields. In Proceedings of the Asian Conference on Computer Vision, Tokyo, Japan, 18–22 November 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 19–34.
47. Laina, I.; Rupprecht, C.; Belagiannis, V.; Tombari, F.; Navab, N. Deeper depth prediction with fully convolutional residual networks. In Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV), Stanford, CA, USA, 25–28 October 2016; pp. 239–248.
48. Xu, D.; Ricci, E.; Ouyang, W.; Wang, X.; Sebe, N. Multi-scale continuous crfs as sequential deep networks for monocular depth estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 5354–5362.
49. Mousnier, A.; Vural, E.; Guillemot, C. Partial light field tomographic reconstruction from a fixed-camera focal stack. *arXiv* **2015**, arXiv:1503.01903.
50. Levin, D. The approximation power of moving least-squares. *Math. Comput. Am. Math. Soc.* **1998**, *67*, 1517–1531. [CrossRef]

51. Salokhiddinov, S.; Lee, S. Depth from focus for 3D reconstruction by iteratively building uniformly focused image set. In *ACM SIGGRAPH 2018 Posters*; ACM: New York, NY, USA, 2018; pp. 1–2.
52. Park, J.; Tai, Y.W.; Cho, D.; So Kweon, I. A unified approach of multi-scale deep and hand-crafted features for defocus estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1736–1745.