

# Composite Focus Measure for High Quality Depth Maps

Parikshit Sakurikar and P. J. Narayanan

Center for Visual Information Technology - Kohli Center on Intelligent Systems,  
International Institute of Information Technology - Hyderabad, India

{parikshit.sakurikar@research., pjn@}iiit.ac.in

## Abstract

*Depth from focus is a highly accessible method to estimate the 3D structure of everyday scenes. Today's DSLR and mobile cameras facilitate the easy capture of multiple focused images of a scene. Focus measures (FMs) that estimate the amount of focus at each pixel form the basis of depth-from-focus methods. Several FMs have been proposed in the past and new ones will emerge in the future, each with their own strengths. We estimate a weighted combination of standard FMs that outperforms others on a wide range of scene types. The resulting composite focus measure consists of FMs that are in consensus with one another but not in chorus. Our two-stage pipeline first estimates fine depth at each pixel using the composite focus measure. A cost-volume propagation step then assigns depths from confident pixels to others. We can generate high quality depth maps using just the top five FMs from our composite focus measure. This is a positive step towards depth estimation of everyday scenes with no special equipment.*

## 1. Introduction

Recovering the 3D structure of the scene from 2D images has been an important pursuit of Computer Vision. The size, relative position and shape of scene objects play an important role in understanding the world around us. The 2.5D depth map is a natural description of scene structure, corresponding to an image from a specific viewpoint. Multi-camera arrangements, structured lights, focus stacks, shading etc., can all recover depth maps under suitable conditions. Users' experience and understanding of the environment around them can be improved significantly if the 3D structure is available. The emergence of Augmented and Virtual Reality (AR/VR) as an effective user interaction medium enhances the importance of easy and inexpensive structure recovery of everyday environments around us.

Depth sensors using structured lights or time-of-flight cameras are common today, with a primary use as game appliances [13]. They can capture dynamic scenes but have

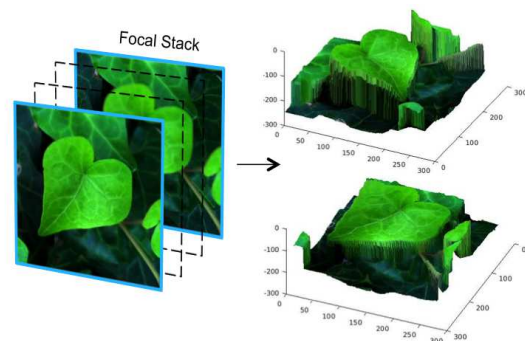


Figure 1. A coarse focal stack of an outdoor scene and its surface-mapped 3D depth is shown from two different viewpoints. The depth-map is computed using our composite focus measure. The smooth depth variation along the midrib of the leaf is clearly visible in the reconstructed depth rendering.

serious environmental, resolution and depth-range limitations. Multi-camera setups are more general, but are unwieldy and/or expensive. Focus and defocus can also provide estimates of scene depth. Today's DSLR cameras and most mobile cameras can capture focal stacks by manipulating the focus distance programmatically. Thus, depth from focus is a promising way to recover 3D structure of static scenes as it is accessible widely.

We present a scheme to recover high quality depth maps of static scenes from a focal stack, improving on previous depth-from-focus (DfF) methods. We show results on several everyday scenes with different depth ranges and scene complexity. Figure 1 is an example of robust depth recovery that we facilitate. The specific contributions of this paper are given below.

1. **Composite Focus Measure:** A focus measure (FM) to evaluate the degree of focus or sharpness at an image pixel is central to DfF. Several focus measures have been used for different scenarios. We combine them into a composite focus measure (cFM) by analyzing their *consensus* and *correlation* with one another over 150 typical focal stacks. The cFM is a weighted combination of individual adhoc FMs with weights com-

puted off-line. In practice, a combination can involve as few as two FMs or as many as all of them.

2. **Depth Estimation and Propagation:** We use a two-stage pipeline for DfF, with the first stage estimating a fine depth at each pixel using a Laplacian fit over the composite focus measure. This gives both a depth estimate and a confidence value for it. In the second stage, a cost-volume propagation step distributes the confident depth values to their neighborhoods using an all-in-focus image as a guide.

We present qualitative and quantitative results on a large number and variety of scenes, especially everyday scenes of interest. The depth maps we compute can be used for applications that RGBD images are used for, typically at resolutions and fidelity higher than them.

## 2. Related Work

**Depth from Focus/Defocus:** The computation of depth from multiple focused images has been explored in the past [2, 4, 20, 29]. Defocus cues have also been used [3, 7, 9, 19, 22, 23, 30] to estimate scene depth. In most methods, depth is estimated from the peak focus slice computed using per-pixel focus measures. Pertuz *et al.* [24] analyze and compare several focus measures independently for DfF. They conclude that Laplacian based operators are best suited under normal imaging conditions. In [20], the Laplacian focus measure is used to compare classical DfF energy minimization with a variational model. A new RDF focus measure was proposed in [28], with a filter shape designed to encode the sharpness around a pixel using both local and non-local terms. Mahmood *et al.* [18] combined three well known focus measures (Tenengrad, Variance and Laplacian Energy) in a genetic programming framework. Boshtayeva *et al.* [4] described anisotropic smoothing over a coarse depth map computed from focal stacks. Suwanakorn *et al.* [29] proposed a joint optimization method to solve the full set of unknowns in the focal stack imaging model. Methods such as [4, 20] can benefit from the composite focus measure we propose in this work.

**Focal Stacks and All-in-focus Imaging:** Focal stacks are images of the scene captured with same camera settings but varying focus distances. Usually a focal stack has each scene point in clear focus in one and only one image. Focal stacks enable the generation of all-in-focus (AiF) images where each pixel corresponds to its sharpest version. Generating the best in-focus image has been the goal for several works [1, 15, 21, 32]. Reconstruction of novel focused images has also been achieved using focal stacks [14, 10, 11, 21, 29, 33].

Focal stacks can be captured without special equipment or expensive cameras. Several mobile devices can be programmed to capture multiple focused images sequentially. Region-based focus stacking has also been used in the past on mobile devices [26]. Most DSLRs can automatically capture focal stacks. MagicLantern [17] provides controls on Canon DSLRs to set focus limits and focus ring movement between consecutive slices. Focal stacks are used for scene depth recovery in DfF methods.

## 3. Composite Focus Measure

Depth from focus (DfF) methods estimate the degree of focus at a pixel by evaluating a focus measure (FM) across the slices of a focal stack. A focus measure is expected to peak at the slice that was focused closest to the true depth of the pixel. The standard idea in DfF is to assign depth based on the peak of the focus measure. The resulting depth maps are usually noisy and constrained in depth resolution to the number of focal slices.

Two factors critically affect good depth estimation: Quality of the FM and its Region of Support. No single focus measure works well in all situations, whether it uses statistical, spectral, gradient, or other properties of the pixel neighborhood. The response of a focus measure depends significantly on the underlying scene structure and intensities. For most focus measures, the size of the region of support plays an important role in the identification of the focus peak. Smaller regions usually have high specificity, but noisy estimates. Larger neighborhoods provide stable estimates but cause dilation across depth edges.

Pertuz *et al.* [24] analyzed 36 different focus measures individually to characterize their sensitivity with respect to support window size, image contrast, noise and saturation. Their analysis provided no definitive recommendation about the best focus measure as different ones exploit different properties and perform well on different scenes. This suggests that a combination of FMs can work well for more varied situations. The key objective of our work is to identify a composite focus measure (cFM) as a weighted combination of the individual FMs. We do so by analyzing the performance of 39 FMs (all from [24], two additional measures which featured later in Boshtayeva *et al.* [4] and the RDF from [28]) in the context of DfF on every pixel of a dataset of about 150 focal stacks.

Selecting the best subset of focus measures from a large number of them is a challenging problem. Supervised approaches with principled learning of weights for a composite focus measure are not feasible, due to the lack of ground truth data. Capturing large number of aligned focal stacks and depth maps can enable supervised learning of FM weights or the use of deep learning methods to directly come up with a robust composite measure. This is a direction we intend to pursue in the future.

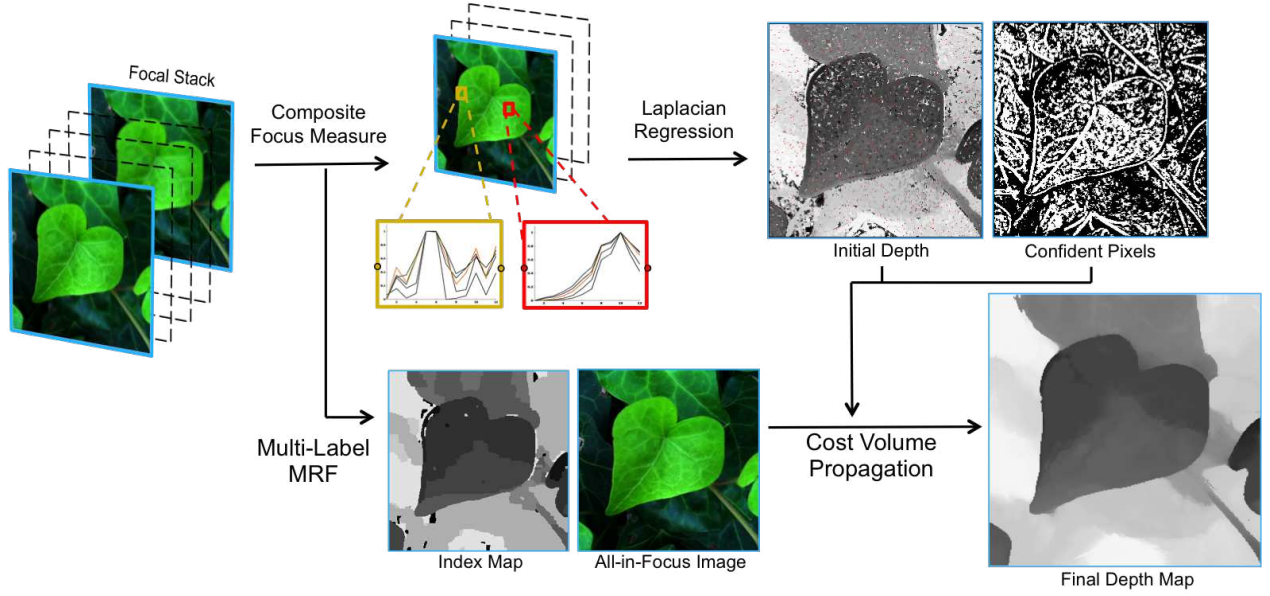


Figure 2. Our proposed pipeline to compute smooth depth-maps from focal stacks. The composite focus measure is evaluated at each pixel of the focal stack and the focus responses are used to (a) generate a high resolution depth value at each pixel using Laplacian regression and (b) generate an all-in-focus image using a multi-label MRF optimization. The all-in-focus image and the confident high resolution depths are used together to compute a smooth depth map using Cost-Volume Filtering.

In the absence of ground truth depth, unsupervised feature selection is the natural candidate for FM selection. Unsupervised methods use unified learning frameworks that simultaneously estimate the structure of the data and the best set of features that describe the data [6, 8]. However, selecting the best combination of focus measures is different from the feature selection problem. In feature selection, the goal is to identify the best subset of representative features which define the data well, and each selected feature usually encodes different information about the underlying data. The selection process thereby maximizes diversity between individual features. For the selection of focus measures, all the features represent the same information - the amount of focus at a pixel. Therefore, the agreement of different focus measures is important.

Traditional methods for unsupervised feature selection of focus measures [6, 8] perform poorly for DfF (Figures 4, 5), as expected. The top-ranked measures according to [6] exhibit different focus peaks at most pixels, since FMs with diverse responses are selected. For DfF, it is important to select those FMs which agree with one another. However, since we use a diverse collection of FMs, some FMs may give near identical responses to others. Measures that agree on the focus peak but not at other slices should ideally be part of the composite focus measure. Thus, we seek *consensus* among the FMs but *not chorus*. In the following sections we describe our strategy to compute the composite focus measure by looking for high-consensus FMs which do not have high correlation.

### 3.1. Consensus of Focus Measures

We start with 39 focus measures reported in the literature [4, 24, 28] and want to identify a small subset that works best for DfF. The consensus or agreement between different FMs on the peak location is a strong indication of the fidelity of each focus measure response. We propose two unique methods to evaluate consensus: Max consensus and MRF consensus. In Max consensus, the focal slice at which most focus measures peak is identified for each pixel. The focus measures that peak within a small neighborhood of this slice are assumed to be in consensus. The  $C_{max}$  function computes max consensus as:

$$C_{max}(F_j; p) = \begin{cases} 1 & \text{if } \arg \max_l F_j(p, l) \\ & \in [m(p) - w, m(p) + w] \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

Here  $m(p)$  is the focal slice at which maximum number of measures peak for pixel  $p$ ,  $F_j(p, l)$  the  $j^{th}$  focus measure response at pixel  $p$  of slice  $l$  and  $w$  denotes a small neighborhood around  $m(p)$ . We choose  $w$  to be 10% of the number of focal slices in the stack. This corresponds to a small depth neighborhood as the focus steps in our focal stacks are uniform.  $w$  can be parameterized based on the blur difference between two slices in case of non-uniform focus steps.

For MRF consensus, we use all focus measures to build a smooth index map for the focal stack using MRF based

energy minimization [5]. The data cost  $D_L(p)$  of labeling a pixel  $p$  to focal slice index  $L$  is computed as the normalized sum of all FM responses at the pixel:

$$D_L(p) = e^{-W}, \quad W = \sum_{j=1}^{n_{FM}} \frac{F_j(p, L)}{\sum_l F_j(p, l)} \quad (2)$$

where  $n_{FM}$  denotes the number of focus measures and  $F_j(p, L)$  is the  $j^{th}$  focus measure at pixel  $p$  for the  $L^{th}$  focal slice. A multilabel Potts term is used to assign smoothness costs.

The result of the MRF optimization is a globally smooth index labeling for each pixel. We define MRF consensus as the agreement of focus measure responses with the MRF labels. The  $C_{mrf}$  function computes the MRF consensus as:

$$C_{mrf}(F_j; p) = \begin{cases} 1 & \text{if } \arg \max_l F_j(p, l) \\ & \in [i(p) - w, i(p) + w] \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

Here  $i(p)$  is the label index assigned by the MRF at pixel  $p$  and other parameters are same as earlier.

The  $C_{max}$  consensus score for an FM indicates the number of times the FM was in agreement with the slice favored by the maximum number of FMs while the  $C_{mrf}$  score indicates its agreement with global focus peak labels. To encode these consensus properties together, we build a cumulative consensus score for each FM as  $C_{max} + C_{mrf}$  across all the pixels of a large data corpus of 150 focal stacks. The FMs are now ranked based on the cumulative consensus score starting with the highest. We represent the FMs in this paper using the naming convention of Pertuz *et al.* [24]; the additional measures are labeled as HFN (Frobenius Norm of the Hessian), DST (Determinant of Structure Tensor) and RDF (Ring Difference Filter).

### 3.2. Correlation of Focus Measures

The list of FMs we use contain near-identical or highly correlated measures. These will naturally be in consensus with each other as they encode very similar information. We would like to choose only one of each highly correlated pair of FMs. To do this, we compute all  $\binom{39}{2}$  pairwise correlation values between the FMs across the 150 focal stacks. The correlation between a pair of measures  $F_i$  and  $F_j$  is defined as

$$Cor(F_i, F_j) = \sum_{FS} \sum_p \sum_l \sqrt{(F_i(p, l) - F_j(p, l))^2}, \quad (4)$$

where FS indicates all focal stacks,  $p$  indicates all the pixels in a focal slice and  $l$  indicates the number of slices in the stack.

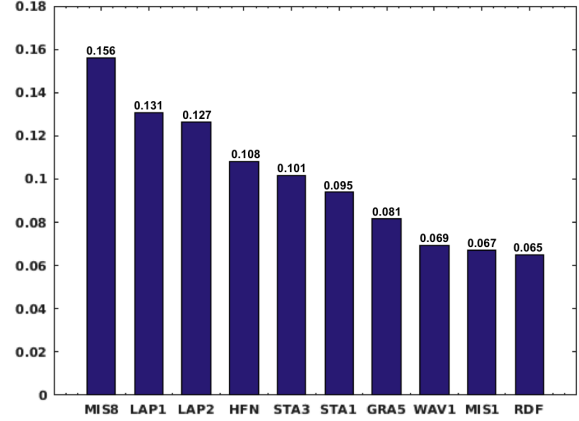


Figure 3. Top 10 focus measures with a high degree of consensus but not high correlation. The normalized consensus score is shown on the Y-axis. This score is used as the weight for creating the composite focus measure.

We now isolate all pairs of FMs which show a correlation greater than 80%. From each of these FM pairs, the FM with the higher consensus score is retained and the other is removed. This process is applied transitively, i.e. if the correlation between A:B and B:C is greater than 80%, then the measure with the highest consensus score is retained (say A) and the other measures (B and C) are removed. On iteratively parsing through all pairs of highly correlated FMs, we arrive at the list in Figure 3, which shows the top ten FMs with high consensus but not high correlation.

A weighted combination of the top five focus measures of Figure 3 forms our composite focus measure (cFM). The weights for each measure are assigned based on their normalized cumulative consensus score. It is interesting to note that well-known and robust FMs from three different focus measure families [24] - laplacians, gradients and variance - are naturally selected for the cFM, along with newer measures such as the HFN.

Using more than five FMs in the cFM results in minor improvements in depth quality at the cost of increased computation, while using lesser FMs results in loss of quality. To test the generalization of the cFM, we also evaluate consensus and correlation measures separately for subsets of the 150 focal stacks. The subsets are based on different scene categories such as texture complexity, amount of blur, position and spread of objects, etc. Our experiments suggest that such categorization has little impact on the ranking of the FMs. The top five FMs remain the same for almost all subsets. Even over uncategorized subsets of the 150 stacks, the top five measures remain the same, suggesting good generalization of the cFM. We now describe how we can use the cFM to build high-quality depth-maps.



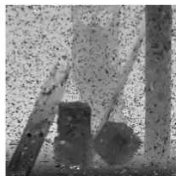
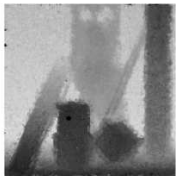
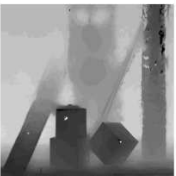
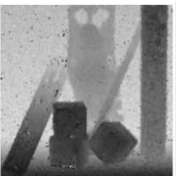










MCFS-5	WAV1	RDF	LAP2	Ours – Top 5	Ground Truth	In-Focus Image
 16.20 dB	 20.31 dB	 21.95 dB	 21.43 dB	 22.37 dB		 Buddha
 22.01 dB	 25.19 dB	 26.64 dB	 27.18 dB	 29.05 dB		 Medieval

Figure 4. Quantitative Evaluation on two synthetic datasets from [31]. We generate 25 focal slices using the ground truth depth map and use our two stage DfF pipeline to compute depth using different FMs. Our composite focus measure performs better than the top single measures from [24, 28], which is visible in the images and reflected in the PSNR (in *dB*) reported below each depth-map. MCFS-5 denotes selecting top five measures using the unsupervised feature selection approach of [6]. We report PSNR to indicate the comparison between 8-bit grayscale ground truth depth maps and high resolution 8-bit depths computed using our method.

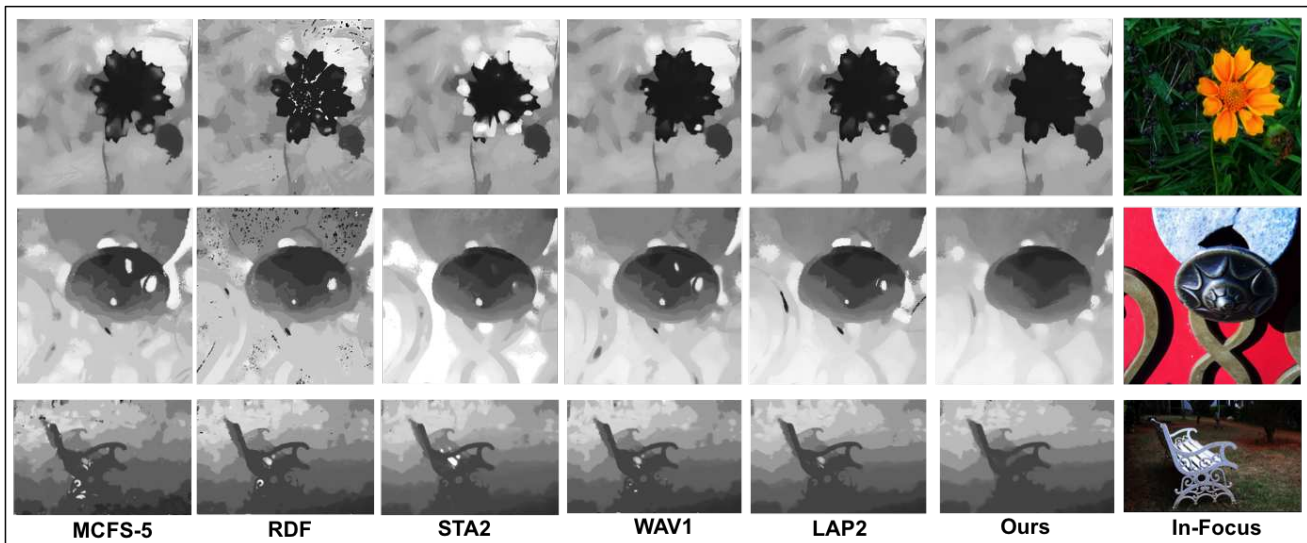


Figure 5. Qualitative comparison of the top individual focus measures from [24], our implementation of [28] and our composite focus measure. Our two-stage DfF pipeline is used in all cases. The composite focus measure captures the true focus profile even at difficult scene locations. MCFS-5 denotes using the top five focus measures selected using the unsupervised approach of [6].

## 4. Depth Estimation and Propagation

Figure 2 shows the pipeline of our depth-from-focus method. We first build a high resolution but noisy depth-map by fitting a Laplacian distribution to the composite focus measure at each pixel. We then build a high-resolution cost volume (256 depth labels) corresponding to the confident depth labels and use an MRF-based in-focus image for guidance to compute a smooth depth map of the scene.

### 4.1. Depth from Laplacian Regression

A Laplacian distribution has been shown to be a good model for depth [27] as it captures sharp depth edges well.

Since the focus profile of a pixel is expected to be closely related to its depth profile, we estimate the depth of a pixel by fitting a non-linear Laplacian distribution over its composite focus measure. For each pixel, we collect the focus responses of the composite focus measure as a set of data points (insets of Figure 2) and fit a Laplacian distribution over them. The Laplacian distribution has the form

$$g(x|\mu, b) = \frac{1}{2b} e^{-\frac{|x-\mu|}{b}}, \quad (5)$$

with  $\mu$  denoting the location and  $b$  denoting the scale or diversity.

We use a standard iterative non-linear regression frame-

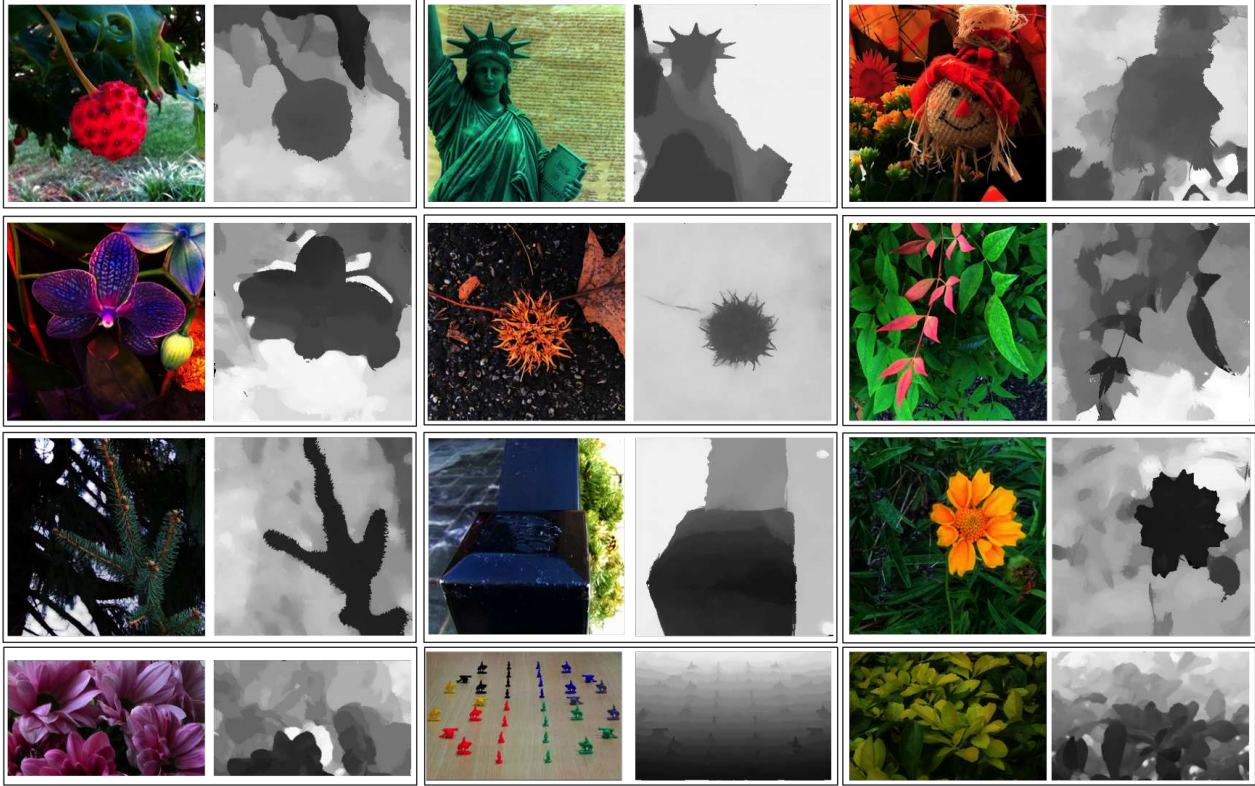


Figure 6. All-in-focus image and computed depth maps for different focal stacks from [16] and focal stacks that we captured. The first three rows show 9 focal stacks from [16] with different focal resolutions, indoor/outdoor scenes and varying levels of scene texture. The last row consists of three focal stacks that we captured using Canon EOS 1100D, 70D and 350D from left to right. These focal stacks had high focal resolution and degree of blur. Our composite focus measure and DfF pipeline clearly produces good depth reconstruction for various scene types.

work for least squares fitting at each pixel. The estimated  $\mu$  represents a smooth depth value. The real-valued  $\mu$  estimates have a much finer resolution than the number of focal slices in the stack. We linearly rescale the values of  $\mu$  from  $[1, L] \subset \mathbb{R}$  to  $[0, 255] \subset \mathbb{Z}$ , representing high resolution depths. This linear scaling can be appropriately adjusted based on the blur between pairs of focal slices if the focal stack was captured with non-uniform focus steps. The rescaled  $\mu$  at each pixel is notated as the initial depth  $D_i(p)$  at the pixel. Laplacian fitting over the composite focus measure is a departure from standard DfF methods which simply assign the focal slice label at which a focus measure peaks. For example, in Figure 2, the focal stack consists of 11 focal slices and the depth resolution reported in several DfF methods is thereby limited to 11 depths, similar to the index map shown in the figure. Our initial depth after Laplacian regression (right-hand side of Figure 2) is already made up of 243 unique depth values.

The scale  $b$  of the Laplacian encodes the confidence of the depth value. Higher the value of  $b$ , lower is the confidence of computed depth. After normalizing the values of  $b$ , the confidence at each pixel is recorded as  $D_c(p) = 1 - b(p)$ .

## 4.2. Cost Volume Propagation

We use the Cost Volume Filtering technique [25] to propagate confident depth labels to other pixels. We build a high resolution cost-volume of 256 volumetric indices, each representing a depth value. The cost of a pixel for every label is assigned based on  $D_i$  and  $D_c$ . High confidence pixels are assigned zero cost to the label corresponding to their depth value  $D_i$ , and linearly increasing costs for other labels. All other pixels are assigned zero costs for all labels.

$$C_i(p) = \begin{cases} |D_i(p) - i| & \text{if } D_c(p) > t \\ 0 & \text{otherwise} \end{cases}. \quad (6)$$

Here,  $C_i(p)$  is the cost of assigning the label  $i$  to pixel  $p$ , with  $i$  indicating the 256 depth labels of the cost volume,  $D_i$  the initial depth and  $D_c$  the confidence from Laplacian regression, and  $t(= 0.85)$  is the empirically computed confidence threshold.

A guided filtering operation over the cost volume generates the labeling for each pixel [12]. Guided image filtering enforces neighbourhood consistency along depth boundaries based on the intensity changes in a guidance image.

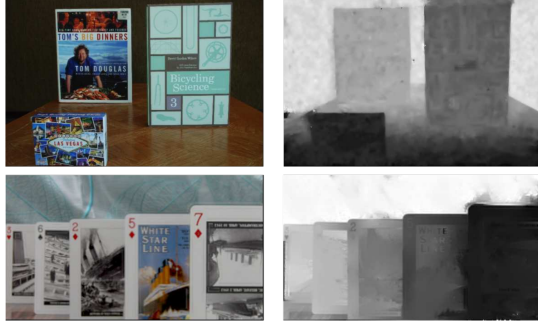


Figure 7. Focal stacks and computed depth-maps for the quantitative comparison of our approach with that of [29].

We generate an all-in-focus image as the guidance image using a multi-label MRF over the composite focus measure. The data term and smoothness costs are assigned similar to Eqn. 2, with the sum of the composite focus measure providing the data cost for each pixel.

After filtering the cost volume  $C_i$  using the guidance image, a smooth depth map can be computed from the filtered cost volume  $C'_i$  in a winner-takes-all manner:

$$\mathcal{D}(p) = \arg \min_i C'_i(p) \quad (7)$$

Figure 2 shows the depth map generated using the guidance image and cost volume propagation.

## 5. Experiments and Results

We demonstrate results on real world focal stacks that we captured as well as other focal stacks used earlier [4, 16, 29]. Our data corpus for computing the cFM consists of 150 focal stacks with varying scene characteristics such as depth range, degree of blur, number of focal slices, textures in the scene, indoor/outdoor illumination etc. We use 100 focal stacks from the light-field saliency dataset [16] representing everyday scenes and having focal resolution from 3 slices to 12. We also use 40 other focal stacks with high degrees of defocus blur. These were captured by us on DSLR cameras such as Canon 70D, 350D, 1100D as well as mobile devices such as the Nexus 5X. These vary in focal resolution from 5 to 40 slices. We also use 10 focal stacks provided by previous researchers [4, 29]. On the Canon DSLR cameras, we used MagicLantern [17] to capture focal stacks and for the Nexus 5X we implemented a custom focal stack capture application using the Android Camera2 API.

We use standard parameter values as defined in [4, 24, 28] for any focus measures that require additional parameters. The offline process of computing the cFM is a compute intensive process. In this step, all FMs are evaluated for three different support-window resolutions of  $3 \times 3$ ,  $7 \times 7$  and  $11 \times 11$  and then averaged, to assemble a cumulative response across multiple regions of support. We reuse

computed numerical values whenever possible, as multiple measures from the same family start with similar numerical computations. All our modules are implemented in Matlab except for the MRF module which is in C++. Once the cFM is computed, the computational complexity of our method is moderate. At runtime, we apply all FMs from the composite focus measure at a window size of  $3 \times 3$  because noisy estimates are acceptable as they average out across the cFM but dilation due to larger window sizes results in more serious depth errors. Applying the composite focus measure, laplacian regression and depth propagation together takes about 60 seconds on a focal stack of  $1k \times 1k$  images on a standard desktop computer. We are building Android and iOS applications which can capture few-sliced focal stacks and generate useful depth maps based on our approach.

We show qualitative and quantitative results to evaluate our method. We compare the effectiveness of our composite focus measure against individual top focus measures defined in [4, 24, 28], using the same two stage DfF pipeline. We perform quantitative evaluation of our depth-map using a few light-field datasets from [31] and also use an evaluation strategy similar to [29]. We provide qualitative comparison with state-of-the-art techniques such as [4, 29] and also demonstrate good quality depth reconstruction on new focal stacks.

### 5.1. Quantitative evaluation

Figure 4 gives quantitative depth reconstruction results for the dataset from [31]. We synthesize 25 focal slices from synthetic light fields (buddha and medieval) and use these focal slices to compute depth using our pipeline. We build a high resolution (256 depths) depth map from just 25 slices, and the depth reconstruction is compared to the available ground truth depth in PSNR (comparing the estimated depth to 8-bit ground-truth grayscale depth). The results show a clear benefit of using our composite focus measure as opposed to other single focus measures. Our composite focus measure also performs much better than the top five (MCFS-5) measures selected from unsupervised feature selection [6].

Figure 7 gives depth computed by our method on two focal stacks. The first is from [29] and the other one is captured by us using a Canon 1100D. In both stacks, the focus ring movement between consequent slices is fixed and thus the depth change between them is quantized. Following [29], known depth values for two objects in the scene are used to compute the depths of the third object. Table 1 gives quantitative comparison of our method with [29]. On our Cards focal stack, we get an RMSE of 0.59 inches for the depth of the cards in the background which are at a depth of more than 30 inches from the camera. Lower error in depth-computation suggests that our method estimates depth maps at a higher quality.



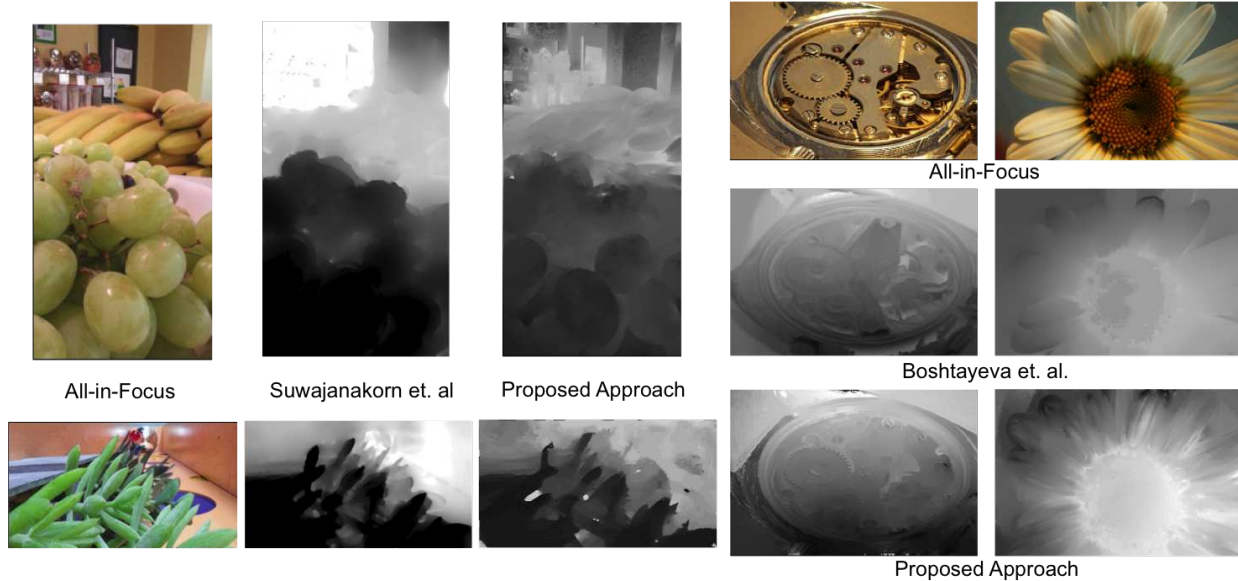


Figure 8. Comparison of our approach with that of Suwajanakorn *et al.* [29] and Boshtayeva *et al.* [4]. The comparison with [29] is shown on the left hand side and with [4] is shown on the right hand side. Our depth maps show improved resolution and smoothness, and the underlying image structure is more precisely retained in the depth image.

Known Depths	Estimated Depth	Ground Truth
$\hat{d}_{box}, \hat{d}_{bike}$	$\hat{d}_{cook} = 27.61$ inches	28 inches
$\hat{d}_{box}, \hat{d}_{cook}$	$\hat{d}_{bike} = 18.64$ inches	18.5 inches
$\hat{d}_{bike}, \hat{d}_{cook}$	$\hat{d}_{box} = 11.83$ inches	12 inches

Table 1. Computed depths for the *Books* focal stack using our method. We observe an average RMSE of 0.45 inches compared to an average RMSE of 2.66 inches reported in [29].

## 5.2. Qualitative results

We demonstrate our results on standard datasets with qualitative comparison to other DfF methods in Figure 8. It can be seen that the detail in the depth map for the fruits dataset and the plants dataset is higher in our results, especially in the regions at low depth values. In the watch dataset, a much smoother variation from near to far can be observed in our results and in the flower dataset, the depth variation in the petals is clearly visible.

Figure 5 shows qualitative performance of our composite focus measure compared to the top individual focus measures from [24, 28] and also over FMs selected using [6]. We also provide depth-maps for focal stacks that we captured and focal stacks that were a part of [16] in Figure 6. The focal stacks shown in Figure 6 have varying degrees of defocus, number of focal slices, depth range, indoor/outdoor illumination conditions etc. The quality of the computed depth-maps indicates that our composite focus measure is robust and provides high quality depth reconstruction.

**Limitations** Our DfF approach is limited to static scenes. Capturing focal stacks of dynamic scenes would require special cameras which can shoot multiple focus distances simultaneously. The assumption that each pixel has a single focus peak can fail if a focal stack ranges from macro to distant objects. Extreme defocus in the foreground can result in previously occluded background pixels appearing sharp, giving two focus peak candidates for some pixel locations. The response of any FM is unreliable at such pixels.

**Dataset** All 150 focal stacks used in our experiments will be made available on our webpage.

**Acknowledgements** This research was partially funded by the TCS Research Scholarship Program.

## 6. Conclusion

In this paper, we demonstrated a novel approach to compute smooth depth-maps from focal stacks. We used the consensus and correlation of 39 different focus measures across a large dataset of focal stacks to identify a weighted combination of FMs as a composite focus measure. The measures with high consensus but not high correlation formed our cFM. The cFM shown in Figure 3 can be used as-is in the future with the normalized scores for each FM. Our two-step depth computation pipeline produces good results on several types of focus stacks ranging from shallow to deep and simple to complex. Our method enables easy and robust capture of 3D scene structure using widely available cameras.



## References

- [1] A. Agarwala, M. Dontcheva, M. Agrawala, S. Drucker, A. Colburn, B. Curless, D. Salesin, and M. Cohen. Interactive digital photomontage. In *ACM Transactions on Graphics*, volume 23, pages 294–302. ACM, 2004. 2
- [2] S. W. Bailey, J. I. Echevarria, B. Bodenheimer, and D. Gutierrez. Fast depth from defocus from focal stacks. *The Visual Computer*, 31(12):1697–1708, 2015. 2
- [3] S. S. Bhasin and S. Chaudhuri. Depth from defocus in presence of partial self occlusion. In *IEEE International Conference on Computer Vision*, volume 1, pages 488–493, 2001. 2
- [4] M. Boshtayeva, D. Hafner, and J. Weickert. A focus fusion framework with anisotropic depth map smoothing. *Pattern Recognition*, 48(11):3310–3323, 2015. 2, 3, 7, 8
- [5] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23, 2001. 4
- [6] D. Cai, C. Zhang, and X. He. Unsupervised feature selection for multi-cluster data. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2010. 3, 5, 7, 8
- [7] S. Chaudhuri and A. N. Rajagopalan. *Depth from defocus: a real aperture imaging approach*. Springer Science & Business Media, 2012. 2
- [8] L. Du and Y.-D. Shen. Unsupervised feature selection with adaptive structure learning. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015. 3
- [9] P. Favaro and S. Soatto. A geometric approach to shape from defocus. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(3):406–417, 2005. 2
- [10] S. W. Hasinoff and K. N. Kutulakos. Light-efficient photography. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(11):2203–2214, 2011. 2
- [11] S. W. Hasinoff, K. N. Kutulakos, F. Durand, and W. T. Freeman. Time-constrained photography. In *IEEE International Conference on Computer Vision*, pages 333–340. IEEE, 2009. 2
- [12] K. He, J. Sun, and X. Tang. Guided image filtering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013. 6
- [13] S. Izadi, D. Kim, O. Hilliges, D. Molyneaux, R. Newcombe, P. Kohli, J. Shotton, S. Hodges, D. Freeman, A. Davison, and A. Fitzgibbon. Kinectfusion: Real-time 3d reconstruction and interaction using a moving depth camera. In *ACM Symposium on User Interface Software and Technology*, 2011. 1
- [14] D. E. Jacobs, J. Baek, and M. Levoy. Focal stack compositing for depth of field control. *Stanford Computer Graphics Laboratory Technical Report*, 1, 2012. 2
- [15] A. Kumar and N. Ahuja. A generative focus measure with application to omnifocus imaging. In *IEEE International Conference on Computational Photography*, pages 1–8. IEEE, 2013. 2
- [16] N. Li, J. Ye, Y. Ji, H. Ling, and J. Yu. Saliency detection on light field. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2014. 6, 7, 8
- [17] Magic lantern. <http://magiclantern.fm/>. 2, 7
- [18] M. T. Mahmood, A. Majid, and T.-S. Choi. Optimal depth estimation by combining focus measures using genetic programming. *Information Sciences*, 181(7):1249–1263, Apr. 2011. 2
- [19] S. Matsui, H. Nagahara, and R. I. Taniguchi. Half-sweep imaging for depth from defocus. In *Advances in Image and Video Technology*, pages 335–347. Springer, 2012. 2
- [20] M. Möller, M. Benning, C.-B. Schönlieb, and D. Cremers. Variational depth from focus reconstruction. *IEEE Transactions on Image Processing*, 24:5369–5378, 2015. 2
- [21] H. Nagahara, S. Kuthirummal, C. Zhou, and S. K. Nayar. Flexible depth of field photography. In *European Conference on Computer Vision*, pages 60–73. 2008. 2
- [22] S. K. Nayar and Y. Nakagawa. Shape from focus. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(8):824–831, 1994. 2
- [23] N. Persch, C. Schroers, S. Setzer, and J. Weickert. Introducing more physics into variational depth-from-defocus. In *German Conference on Pattern Recognition*, pages 15–27, 2014. 2
- [24] S. Pertuz, D. Puig, and M. A. Garcia. Analysis of focus measure operators for shape-from-focus. *Pattern Recognition*, 46(5):1415 – 1432, 2013. 2, 3, 4, 5, 7, 8
- [25] C. Rhemann, A. Hosni, M. Bleyer, C. Rother, and M. Gelautz. Fast cost-volume filtering for visual correspondence and beyond. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2011. 6
- [26] P. Sakurikar and P. J. Narayanan. Dense view interpolation on mobile devices using focal stacks. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 138–143, June 2014. 2
- [27] A. Saxena, S. H. Chung, and A. Y. Ng. Learning depth from single monocular images. In *Advances in Neural Information Processing Systems*, pages 1161–1168, 2005. 5
- [28] J. Surh, H. G. Jeon, Y. Park, S. Im, H. Ha, and I. S. Kweon. Noise robust depth from focus using a ring difference filter. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2, 3, 5, 7, 8
- [29] S. Suwajanakorn, C. Hernandez, and S. M. Seitz. Depth from focus with your mobile phone. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2015. 2, 7, 8
- [30] H. Tang, S. Cohen, B. Price, S. Schiller, and K. N. Kutulakos. Depth from defocus in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2
- [31] S. Wanner, S. Meister, and B. Goldluecke. Datasets and benchmarks for densely sampled 4d light fields. In *Proceedings of the Vision, Modeling, and Visualization Workshop*, 2013. 5, 7
- [32] N. Xu, K.-H. Tan, H. Arora, and N. Ahuja. Generating omnifocus images using graph cuts and a new focus measure. In *International Conference on Pattern Recognition*, pages 697–700, 2004. 2
- [33] C. Zhou, D. Miao, and S. K. Nayar. Focal sweep camera for space-time refocusing. *Technical Report, Department of Computer Science, Columbia University*, CUCS-021-12, 2012. 2