UNSUPERVISED SIMULTANEOUS DEPTH-FROM-DEFOCUS AND DEPTH-FROM-FOCUS

Anonymous authors
Paper under double-blind review

ABSTRACT

If the accuracy of depth estimation from a single RGB image could be improved it would be possible to eliminate the need for expensive and bulky depth sensing hardware. The majority of efforts toward this end have been focused on utilizing geometric constraints, image sequences, or stereo image pairs with the help of a deep neural network. In this work, we propose a framework for simultaneous depth estimation from a single image and image focal stacks using depth-from-defocus and depth-from-focus algorithms. The proposed network is able to learn optimal depth mapping from the information contained in the blurring of a single image, generate a simulated image focal stack and all-in-focus image, and train a depth estimator from an image focal stack. As there is no large dataset specifically designed for our problem, we first learned on a synthetic indoor dataset: NYUv2. Then we compare the performance by comparing with other existing methods on DSLR dataset. Finally, we collected our own dataset using a DSLR and further verify on it. Experiments demonstrate that our system is able to provide comparable results compared with other state-of-the-art methods.

1 INTRODUCTION

Estimation of depth is critical to recover our 3D world and understand it. The capabilities of most computer vision applications are limited by the shortcomings (size, speed, accuracy, and expense) of depth estimating hardware and software e.g. scene understanding, augmented reality and robotics. Conventional depth estimation methods exploit 3D geometric constraints to learn depth information via structure-from-motion (Sweeney et al. (2015) Agarwal et al. (2009) Agarwal et al. (2011)), image sequences or video (Zhou et al. (2017) Babu et al. (2018)), stereo image pairs (Garg et al. (2016) Godard et al. (2017) Poggi et al. (2018)) and structured light cameras (Ryan Fanello et al. (2016)). However, all of these approaches have limitations (e.g. resolution, texture, lighting). SfM has scale ambiguity when using image sequences or video, and there is a camera calibration problem and a lack of compatibility when converting between predicted disparity and real depth values across different scenes. Lastly, these techniques ignore the possible defocus blur at different focal planes based on depth of field and camera settings.

Depth from focus and depth from defocus are two techniques that have been explored in prior art, but their accuracy and performance have not been competitive with the previously mentioned alternative approaches. However, there is strong evidence that multiple biological systems have optimized this approach with great success, e.g. jumping spiders can leap and catch an airborne fly (Nagata et al. (2012) Guo et al. (2019)). This demonstrates the unrealized potential for an artificial solution to leverage the information contained within the differential focusing of objects within a scene. In this paper we present a proposed learning framework, the first unsupervised end-to-end simultaneous training of depth-from-focus and depth-from-defocus networks we know of. Once trained either of these two networks is independently capable of conducting depth estimation either from a single image or focal stack as shown in Fig. 1. This has enabled us to explore the influence of defocus blur in helping depth predictions in an unsupervised manner, instead of directly inferring the depth information from the features and location of each object such as a road, a car, or a person. To evaluate the performance of the proposed framework, we have conducted experiments on relevant benchmarks, both synthetic and real defocused datasets. We have also made comparisons to other recent unsupervised and supervised approaches. Overall, the key contributions of this work are:

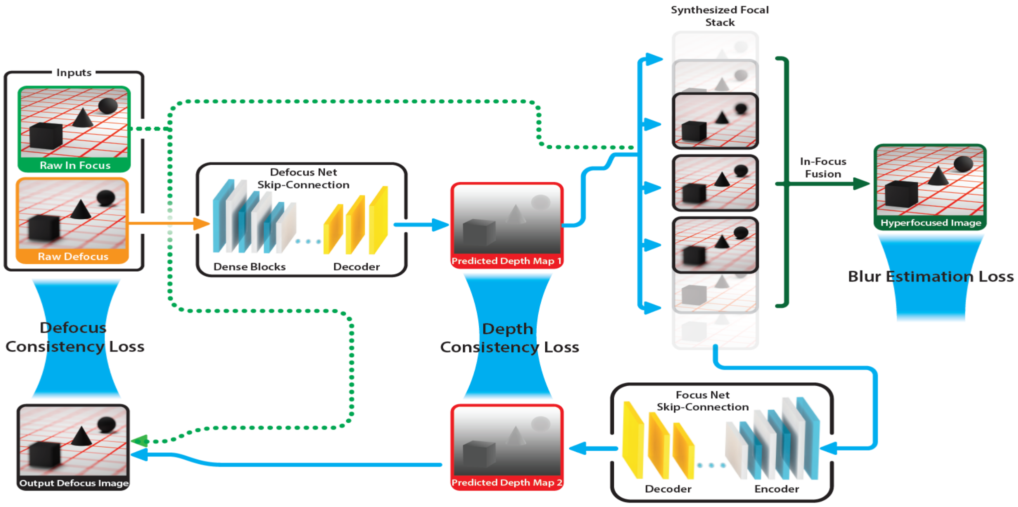

Figure 1: Overview of the proposed framework for simultaneous training of depth-from-defocus and depth-from-focus. Besides depth estimation from two approaches, all-in-focus image is also generated from the network with similar effect as light field cameras.

1) The first end-to-end learning architecture to simultaneously train two networks (FocusNet and DefocusNet) from depth-from-focus and depth-from-defocus algorithms, respectively. 2) By constraining the consistency between the predicted depth maps from a single RGB image or focal stack, we are the first to propose an unsupervised machine learning scheme to estimate depths from focus and defocus. 3) By designing a loss function which measures the degree of focus throughout the image, we have developed a method to determine whether the estimated depth maps are accurate based on the ability to produce an all-in-focus image.

2 RELATED WORK

Supervised Monocular Depth Estimation is capable of producing depth maps from single images (Eigen & Fergus (2015) Liu et al. (2015) Cao et al. (2017) Jung et al. (2017) Ummenhofer et al. (2017)). Eigen et al. (2014) produced depth maps by deploying networks capable of detecting global and textured features using the AlexNet structure (Krizhevsky et al. (2012)). Liu et al. (2015) considered the continuity of the depth values and treated depth estimation as a continuous conditional random field (CRF) learning problem. Cao et al. (2017) took this concept further by formulating depth estimation as a pixelwise classification task, using conditional random field (CRF) as a post-processing scheme. More recent work Ummenhofer et al. (2017) trained an end-to-end network to compute scene depth and camera motion from successive, sequential, and unconstrained image pairs given known optical flow. All such methods require massive labelled images, which is always expensive for many applications.

Unsupervised Monocular Depth Estimation is also capable of producing depth maps from single images, without prior supervisions (Zhou et al. (2017) Yin & Shi (2018) Zou et al. (2018) Garg et al. (2016) Godard et al. (2017) Godard et al. (2019)). Garg et al. (2016) proposed a stereopsis based auto-encoder for learning; Meanwhile, the trained network is capable of producing depth map from a single image. Godard et al. (2017) built upon this work by considering left-and-right pixel disparity consistency loss. Unlike exploiting geometrical cues from stereo pairs, (Zhou et al. (2017) Yin & Shi (2018) Zou et al. (2018)) all succeeded to produce monocular depth estimation methods utilizing camera pose regression, unlabeled video sequences, and photometric consistency between source and target views. All the methods above require stereo image pairs and/or monocular video sequences during the training process, and ignore the blur information from the camera. Additionally, the absolute scales trained with monocular cameras remain ambiguous.

Depth From Defocus/Depth From Focus Distinct from the aforementioned depth estimation methods, depth estimation from defocus reconstructs a pixel-accurate depth map by utilizing blur information to determine the degree by which the corresponding objects deviate from a focal plane. One such approach estimated the spatially varying defocus blur from the ratio of input gradients and re-blurred images (Zhuo & Sim (2011)). Improvements in defocus blur measurements have been realized with coded aperture cameras (Zhou et al. (2009) Ranftl et al. (2016)). Depth From Focus is similar to depth from defocus with the key difference of depth from focus requiring dynamically changing camera parameters during the estimation process. Depth from focus has managed to produce depth maps utilizing focus stacks input from either a set of frames (Suwajanakorn et al. (2015)) or a light field camera (Lin et al. (2015)).

Deep Neural Network and Depth From Focus/Defocus have been individually combined in recent years to increase the accuracy; however, this field of inquiry is still in its infancy. Depth from focus/defocus can potentially alleviate scale ambiguity issues that arise from monocular depth estimation techniques, but existing depth from focus/defocus methods based on deep neural networks mainly rely on supervised learning schemes. Modern filtering methods are capable of producing realistic defocus blur effects on readily available datasets (e.g. NYU or KITTI) for use in depth prediction. A new lightfield dataset has been proposed in Srinivasan et al. (2017) and monocular depth has been predicted from rendered focused images with a deep regression model. However, this approach estimates depth maps from all-in-focus images, which is often not representative of real world cameras. Carvalho et al. (2018) showed that out-of-focus blur improves depth estimation performance; however, this technique relies on the ground truth depth for supervision. Hazirbas et al. (2018) proposed an auto-encoder-style convolutional neural network to estimate depth from a focal stack and ground truth depth map using 4D light field images. In addition to the need for ground truth data, their collected focal stack has small variations in defocus blur, which provides limited blur information for training. To overcome these shortcomings, this paper describes an approach which employs two deep learning networks, one for depth from defocus (DefocusNet) and one for depth from focus (FocusNet). To our best knowledge, this is the first unsupervised learning method which simultaneously trains depth from defocus and depth from focus networks, which can be independently utilized post-training.

3 Simultaneous Depth from Defocus and Focus

In this section, we first describe the principle and method to simulate defocused images. Then we introduce the method applied for all-in-focus image completion. Furthermore, the model architecture of our DefocusNet and FocusNet is explained. Finally, training constraints are provided.

3.1 Thin lens illustration for defocus image generation

Depth from defocus/focus methods are based on the thin lens model illustrated in Fig. 2. f is the focal length. v is the object distance, and d is the scene depth. An object is regarded as "in focus" if it lies in the depth-of-field (DOF) for the camera. Objects outside this range would be regarded as "out of focus." The divergence of the light rays caused by unfocused objects leads to the "Circle-of-Confusion" i.e. blur diameter, which is indicated as ε in Fig. 2. D is the lens diameter, the distance between the sensor and the lens is s. The defocus blur can be expressed with the following geometric relationship:

$$\varepsilon = Ds \cdot |\frac{1}{f} - \frac{1}{v} - \frac{1}{s}| \tag{1}$$

Given a scene with radiance L, the defocus can be expressed by the convolution operation $\overline{L} = L * B_\varepsilon$, B_ε represents the 2D point-spread function parameterized by the blur diameter ε, which is a function of object distance, focal length, and aperture diameter from Eq. 1. By adopting the approaches of Hasinoff & Kutulakos (2007), an approximate layered image formation model with occlusion is expressed as:

$$\overline{L'_k} = \sum_k [(A_k L' + A^*_k {L'_k}^*) * B_\varepsilon(k)] \cdot M_k \tag{2}$$

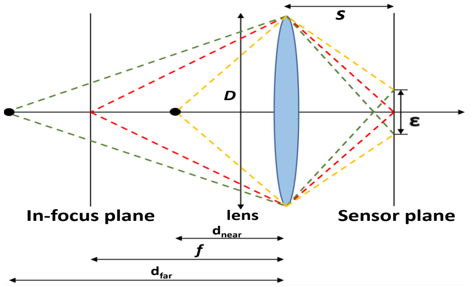

Figure 2: The thin lens model.

where L' is the in-focus radiance or image, the A_k mask corresponds to the objects at the depth k, $A^*_k L'^*$ acts as extended versions of the unoccluded radiance for each layer, and M_k represents the cumulative occlusion from previous defocused layers.

$$M_k = \prod_{k'=k+1}^{K} (1 - A_k * B_\varepsilon(k)) \tag{3}$$

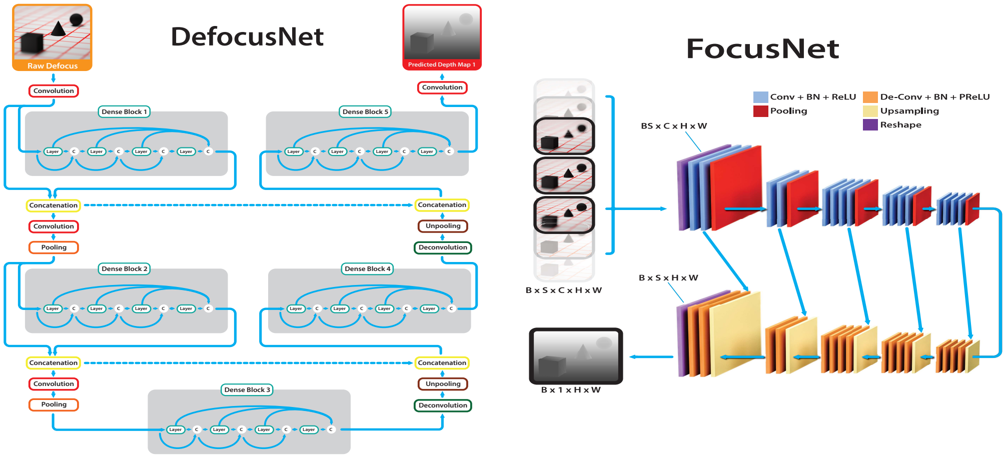

Figure 3: The DefocusNet and FocusNet structures.

This strategy enables the generation of out-of-focus images with variable depths of field (i.e. a synthetic focal stack) by extracting objects at multiple depth k and controlling the position of in-focus plane. Therefore, the influence of blur can be studied without the need of a specialized data set. Note that once real defocus images are able to be captured or collected directly, the generation step can be optionally skipped.

3.2 DefocusNet

To estimate the depth from defocus blur in a image, we apply the densely connected network (DenseNet) Huang et al. (2017) based encoder-decoder architecture refereed as DefocusNet. The DefocusNet aims to estimate a depth map from a defocused image as input. The primary components are: convolutional layers, dense blocks, and transition transformation (transition down and up). In our encoder part, each dense block is composed of a Batch-Normalization layer, Rectified Linear Unit (ReLU) layer and $3 * 3$ convolutional layer, and a Transition Down layer follows each dense block to reduce the feature size. In the decoder, dense blocks are followed by a Transition Up layer to realize an up-sampling operation on the previous feature maps, and concatenate them together with the help of a skip connection to output a predicted depth map.

The motivation of using denseblock for inferring depth information from images with defocus blur is that each layer of this structure is connected to every other previous layer. For each concatenation layer, both the current block and the previous block will be feed-forward and fused together, which is helpful for depth estimation problem, as the features from the previous input can be reused and are capable to enhance the feature propagation. Especially for our depth estimation problem using defocus clues, the depth information is hard to extract and learn from the defocus blur in different degrees. Different from the original DenseNet structure consisting of four dense blocks for feature extraction and one block including the fully connected layer for the target image classification, we modify the original network to make it suitable for single image depth estimation related problem. First we reduce the four dense block to two in the encoding part to achieve a balance of high-capacity network parameters and the efficiency. Then we replace the last block for image classification composed of fully connected layer with up-projection layers to achieve a decoder structure.

3.3 FocusNet

FocusNet accepts the generated focal stack as an input and estimates a separate depth map. Building on top of the VGG-19 network Simonyan & Zisserman (2014), it includes 16 convolutional layers, 5 polling layers and 3 fully connected layers. By removing all the fully-connected layers, adding two extra convolutional layers, and adding a multi-layer concatenation connection, the network is able to preserve more edge information in the estimated depth map. The details for the multi-layer concatenation connection layer is illustrated as follows:

To input the image stack into the $B \times S \times C \times \ H \times W$ (batch size, focal stack, channel, height and width respectively), the focal stack dimension is first embedded together with the batch size to be $BS \times C \times \ H \times W$, and the output from the network is a one-channel depth map in the stack

$BS \times 1 \times \ H \times W$. By reshaping it to $B \times S \times \ H \times W$ and applying a 1×1 convolutional layer, the output dimension is reduced to $B \times 1 \times \ H \times W$, which corresponds to one estimated depth per input focal stack. Followed the FocusNet, the reconstructed defocus image is able to be inferred from the estimated depth map together with the in-focus input. By giving a constraint to make the input defocus image to be consistent with the reconstructed defocus image, the consistency loss is:

$$L_{re} = \frac{1}{N}\sum_{ij}\frac{a}{2}(1 - SSIM(J_{ij}, \tilde{J}_{ij})) + (1-a)\tilde{p}(||J_{ij} - \tilde{J}_{ij}||_1) \tag{4}$$

where the input defocused image is expressed by J, and the reconstructed defocused image is presented by $\tilde{J}$, $\|\cdot\|_1$ represents L_1 norm operator that calculates the mean absolute value. The value for SSIM is from 0 and 1, where 1 indicates a perfect matching. α is a constant parameter and here we choose it as 0.85. To make it more robust, we apply the generalized Charbonnier factor Sun et al. (2010) $\tilde{p}$ to enjoy both benefits of L2 and L1 term.

The similar constraint inspired from Godard et al. (2017) is also applied to the predicted depth maps from the FocusNet and DefocusNet. They are further constrained to be consistent by using the following losses L_{consis} to achieve the self-supervised condition, which is explained as follows:

$$L_{consis} = \frac{1}{N}\sum_{ij}\frac{a}{2}(1 - SSIM(\tilde{D}_{1ij}, \tilde{D}_{2ij})) + (1-a)\tilde{p}(||\tilde{D}_{1ij} - \tilde{D}_{2ij}||_1) \tag{5}$$

where $\tilde{D}_1$ represents the depth map produced from the DefocusNet and $\tilde{D}_2$ is the output estimation from the FocusNet.

3.4 All-in-Focus Image Completion

An all-in-focus image is generated by collapsing the in-focus regions in each image of the synthesized focal stack. First, the same Gaussian blur is applied to each image in the focal stack to smooth the images, then Laplacian of Gaussian (LoG) is utilized to measure the 2nd derivative to obtain the corresponding edge map. As images with higher response contain more sharp edges, by selecting the pixels with the highest edge response following a deblurring kernel related to depth, an all-in-focus image is generated with the same size as the original input and makes the original in-focus region to be more clear. In Fig. 10, the performance of the image completion method is verified on near-focus and far-focus examples, which demonstrates the effectiveness of producing an all-in-focus image by collapsing a focal stack. The blur estimation loss constrains the reconstructed all-in-focus image to be as clear as possible, which will feedback to the estimated depth from depth-from-defocus. Once the depth estimation is accurate, the in-focus region is deblurred correctly. On the other side, if the estimated depth is wrong, the kernel will result in a more blur region. Therefore, blur estimation loss tightly constrains the depth estimation accuracy. More specifically, first a Laplacian kernel is applied and the variance of the response is calculated. If it presents a high variance value, the edges should be more clear. Similarly, if the variance value is low, then the edges and image are regarded to be blur. By setting a range from 0-1000 and normalize the variance response to be in the range of 0 and 1, the blur estimation loss can be designed as the sum of the log loss as:

$$L_{blur} = -\frac{1}{N}\sum_{c=1}^{N}\beta log(n_c) = -\frac{1}{N}\sum_{c=1}^{N}\beta log(\frac{\sum_i\sum_j(\nabla^2 X(i,j)))^2}{M} - \mu^2) \tag{6}$$

where the N is the number of total images. n_c is the normalized coefficient from each image as defined above to evaluate the level of the possible blurring. M is the number of total pixels in an image and μ is the mean value in the image. $X(i,j)$ represents the image pixel at (i,j) in each image. ∇^2 is the Laplacian filter operated on X. β is a constant and here is set to be 2.5. To further prevent the drastic depth change in homogeneous regions and low-texture areas in both DefocusNet and FocusNet, we incorporate a smoothness prior to regularize both estimated depth maps from the DefocusNet and FocusNet inspired by Zhao et al. (2015) and Godard et al. (2017). The edge-aware smoothness loss is stated as follows:

$$L_{smooth} = \sum_{ij}(e^{-\nabla E_{ij}} \cdot \nabla D_{ij}))^2 \tag{7}$$

where ∇ is the first derivative along the spatial directions x and y. It ensures that estimated disparity D_{ij} is guided by the edges of the image E_{ij}. To train the full framework, we rely on the comprehensive objective losses consisting in the weighted sum of the terms: $L_{total} = \lambda_1 L_{re} + \lambda_2 L_{consis} + \lambda_3 L_{blur} + \lambda_4 L_{smooth}$, and the weights here are $\lambda_1 = \lambda_2 = 1.0$, $\lambda_3 = \lambda_4 = 0.2$.

4 Experiments

In this section, we conduct the experiments that assess the performance of our Defocus-Net and Focus-Net in producing estimated depth map and all-in-focus image.

4.1 Training configuration and dataset

The network is implemented with PyTorch and trained using Adam optimizer (Kingma & Ba (2014)) with a learning rate of 0.0001 with a batch size 2. The system is trained jointly on Nvidia Tesla P40 GPU with 24GB cuda memory. A color augmentation is implemented with a 50% chance of random gamma, brightness, and color shifting in the range of [0.8, 1.2], [0.5, 1.5] and [0.8, 1.2] respectively.

NYUv2 RGBD datset Silberman et al. (2012) is comprised of video sequences from a variety of indoor scenes recorded by both the RGB and Depth cameras with more than 120k indoor image pairs. Among them the number of densely labeled pairs of aligned RGB and depth images is 1449. We choose the commonly-used 654 test images from the 1449 labeled RGB-D images to test the performance of our method compared with other methods. To simultaneously train our unsupervised framework composed of FocusNet and DefocusNet, we create a synthetic defocused dataset from the real NYUv2 images based on the provided depth map and raw images, as discussed in Section 3.1. **DSLR dataset** (Carvalho et al. (2018)) contains 110 images and ground truth depths from indoor scenes, with 81 images for training and 29 images for testing, and 34 images from outdoor scenes without ground truth depth. Each scene is acquired with two camera apertures: N = 2.8 as out-of-focus setting and N = 8 as in-focus setting. Because of its limited amount of data, it acts as a supplement real-world split after training on NYUv2.

Our collected dataset consists of eight different indoor scenes including research labs, offices, coffee rooms, meeting rooms, canteens, auditorium halls, library scenes and corridors. Each scene is obtained with two settings of in-focus planes: 1m and 5m respectively. The images are captured by a Nikon D3500 camera but not to it. The first four scenes have 100 images for each and the rest scenes are made up of 500 images for each (totally 2400 images). We randomly select 30 images for the first four scenes and 80 for the other four scenes as testing split. Each scene is put markers on it and acts as the ground truth to test the accuracy in each scene. We put 10 ArUco markers on each testing image to get the depth values between the camera and each marker as the ground truth for testing, as shown in Fig. 4. For a fair comparison, we do not train or fine-tune on it and directly test our models and other recent methods on it (as all other methods for comparison requires perfect-aligned ground truth map). Compared with the DSLR dataset, our collected dataset contains a large number of images (2400 v.s. 110). Compared with the synthetic NYU v2 dataset used in this work, our new dataset is captured from real world indoor environment, which is more realistic and reliable to test the practicability for real applications.

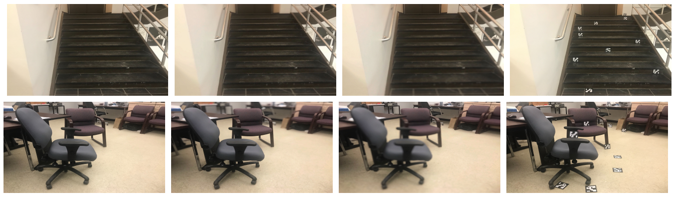

Figure 4: Examples of our collected dataset with different defocus blur and ten markers on it as ground truth.

4.2 Evaluation

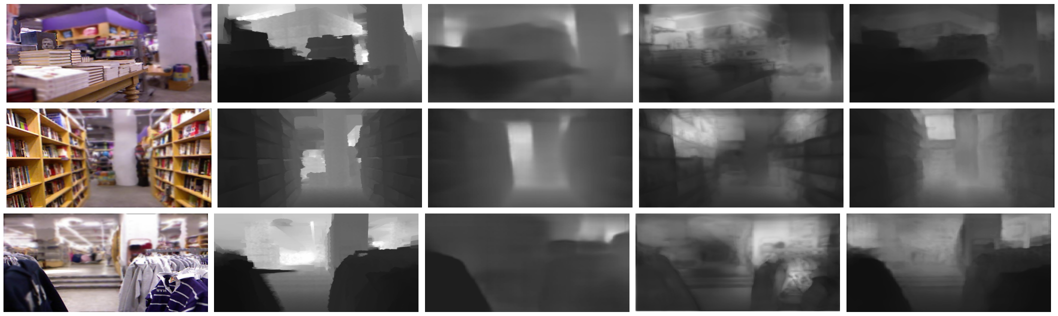

Figure 5: Visual comparison between our method (trained on the synthetic dataset) and other recent single-view depth estimation methods (fine-tuned on the synthetic dataset). Left to right: Input defocus image at the focal plane of 1 meter; Ground truth depth map; Predicted depth map from Laina et al. (2016); Predicted depth map from Alhashim & Wonka (2018); Our output from DefocusNet. Brighter color represents a farther distance.

In this section, we demonstrate the results trained on the NYUv2 dataset with synthetic defocus blur at different distances. We perform two experiments: first we compare with other recent methods on estimating depth map of a single defocus image to demonstrate the effectiveness of the DefocusNet on extracting blur information and estimating depth map from defocus image. In Fig. 5, visual effect of our DefocusNet and other methods for comparison is presented. From the result, it shows that our proposed framework for estimating depth from one single defocus image has a better performance in estimating depth values in blurred regions. Compared with Laina et al. (2016), our method can prevent unexpected ghosting and over-smoothness issues. Compared with Alhashim & Wonka (2018), our model can prevent large holes and severe discontinuities in predicted depth maps.

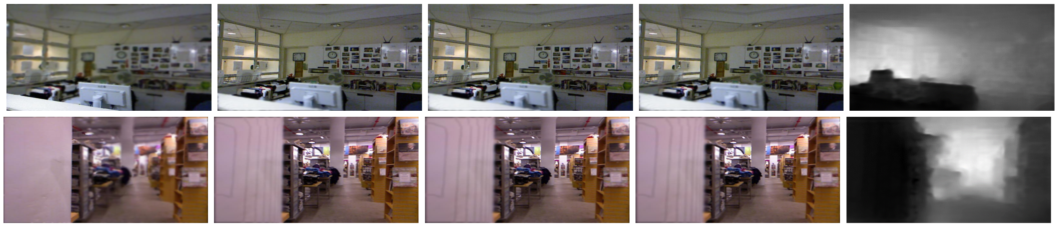

Figure 6: Visual performance of our Focus-Net depth estimation method. From left to right: focal stack at different focal plane: 1m, 3m, 5m, 9m, and estimated depth map from Focus-Net.

To verify the effectiveness of our depth from focus model from the synthetic NYUv2 dataset, we input the focal stack which simulates focus at different focal plane setting as shown in the Fig. 6. It can be observed that with multiple images at different focal planes as input, the FocusNet is capable to learn a relation between the depth map and multiple images of different degrees of defocus blur, and predict an accurate result. As we can observe in Fig. 6, by feeding images of different blurs from near to far field, the estimated depth is able to recover the blur and extract depth information from the image blur as expected.

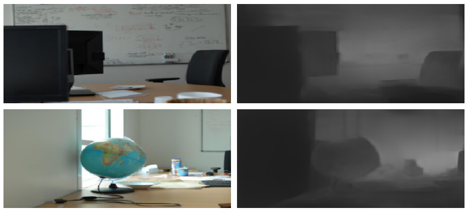

Figure 7: Further verification on realistic scenes from DSLR dataset. First column: Input real defocus images; Second column: Predicted depth maps from DefocusNet. Our model can also work well in real scenes with real defocus blur.

In Table. 1, quantitative comparison against other recent methods on the synthetic dataset and ablation analysis are conducted. All compared monocular methods are trained with direct supervision, and then fine-tuned on the synthetic dataset. It can be observed that our method achieves a higher accuracy than the previous methods for defocused image. As single precision is not sufficient to evaluate in case of Root Mean Squared (RMS) error and Relative (Rel) error. We measure both error (RMS, Rel) and accuracy metrics (δ_1, δ_2, δ_3) to perform a better evaluation. Fig. 5, Fig. 6 and Table. 1 together reflect the superior visual and quantitative performance of our framework (DefocusNet and FocusNet) on the synthetic NYUv2 dataset compared with other recent methods.

Method	$\delta1$	$\delta2$	$\delta3$	RMS	Rel
Liu et al. (2016)	0.652	0.763	0.913	0.997	0.273
Moeller et al. (2015)	0.670	0.778	0.912	0.985	0.263
Suwajanakorn et al. (2015)	0.688	0.802	0.917	0.950	0.250
Eigen et al. (2014)	0.662	0.773	0.910	0.987	0.268
Laina et al. (2016)	0.693	0.862	0.937	0.761	0.192
Xu et al. (2017)	0.698	0.872	0.937	0.768	0.179
Alhashim & Wonka (2018)	0.719	0.875	0.948	0.637	0.172
Lee et al. (2018)	0.701	0.879	0.946	0.723	0.181
Wofk et al. (2019)	0.667	0.851	0.929	0.972	0.226
Gur & Wolf (2019)	0.720	0.887	0.951	0.649	0.184
Ours Defocus-Net w/o smooth	0.729	0.886	0.950	0.628	0.176
Ours Defocus-Net full	0.732	0.887	0.951	0.623	0.176
Ours Defocus-Net gt	0.921	0.989	0.996	0.372	0.084
Our Focus-Net w/o smooth	0.740	0.889	0.946	0.619	0.173
Our Focus-Net full	0.748	0.892	0.949	0.611	0.172
Our Focus-Net gt	0.936	0.990	0.998	0.328	0.075

Table 1: Depth prediction result compared with other methods fine-tuned with defocus images as input on the synthetic NYUv2 dataset.

However, to test the effectiveness of the proposed method in real-world scenes, we need to validate our method on the real defocused images. Fig. 7 shows the visual performance on the depth estimation output from real defocus images (N=2.8). The results show that through training on a large synthetic dataset from Sec. 3.1, the models still can infer reasonable outputs without further training on it. To get a fair comparison on DSLR testing split to get the quantitative results, we further fine-tune both our trained model and those methods for comparison above on 81 training images with ground truth depth. It can be observed from Table. 2, limited by a small amount of images in DSLR dataset, the predicted results for most of the methods suffer a slight drop compared with the reported results on synthetic NYUv2 dataset. But our method still performs well in real-world images with defocus blur compared with other approaches.

Methods	$\delta1$	$\delta2$	$\delta3$	RMS	Rel
Laina et al. (2016)	0.679	0.857	0.932	0.734	0.194
Xu et al. (2017)	0.680	0.859	0.937	0.711	0.194
Alhashim & Wonka (2018)	0.702	0.874	0.940	0.658	0.186
Lee et al. (2018)	0.683	0.862	0.934	0.731	0.190
Wofk et al. (2019)	0.653	0.841	0.926	0.892	0.203
Ours	**0.726**	**0.883**	**0.941**	**0.629**	**0.179**

Table 2: Depth prediction comparison with defocus image as input on real-world DSLR dataset.

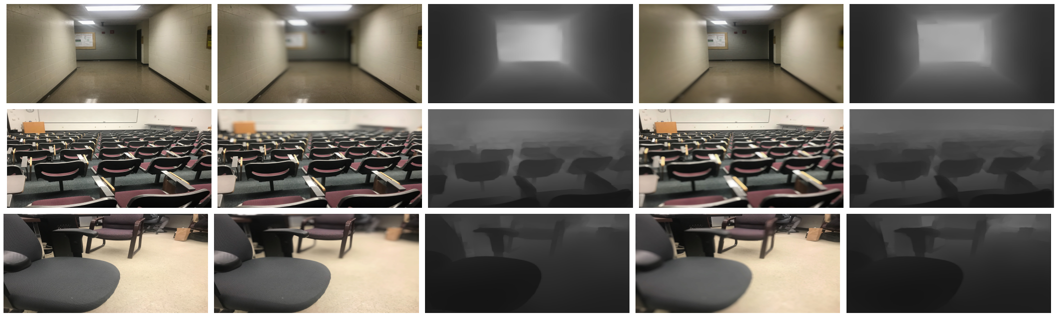

Figure 8: Depth estimation from our collected dataset with 1m and 5m focus. First column: raw in-focus images; Second column: images focusing on 1m; Third column: estimated depth for images focusing at 1m; Fourth column: images focusing at 5m; Fifth column: estimated depth for images focusing at 5m.

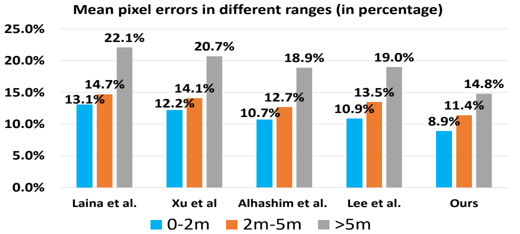

Figure 9: Percentage of mean pixel errors in different distance range.

Due to the limited number of images in the DSLR dataset to evaluate the performance in real scenarios, we evaluate the visual performance and quantitative comparison on the collected dataset, which contains 440 images for testing. To get a fair comparison, we do not conduct fine-tuning or re-training on it, and directly apply the trained model on synthetic NYUv2 dataset to test. From Fig. 8, it can be observed that our method is capable to recover the defocus blur from near-focus and far-focus images, and predict a high-quality depth maps from it, though suffering from slight ghost on the blurred regions of the input images. Fig. 9 provides a comparison of mean pixel errors in percentage with depth values from added markers in different ranges (0-2m, 2-5m and larger than 5m). We can observe that in the range between 0-2m, our method is 2% over the second best result in this range and 4.2% over the worst one in this range. And in the range of >5m, our method is 4.1% over the second best one and 7.3% improvement on top of the worse one in this range. Judging from the percentage and actual numbers, the ambiguity of depth and mean errors of depth increase as the increasing of the distance.

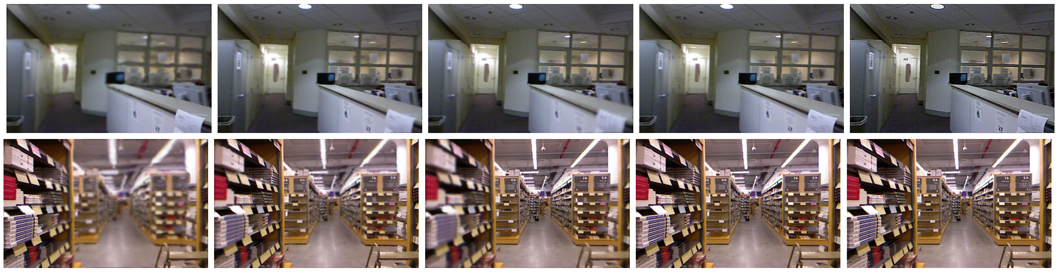

Figure 10: Visual result of our proposed hyper-spectral fusion to get all-clear image at different focal planes. Left to right: three images from different focal stacks, raw input, our hyper-spectral image.

Finally, the performance of our all-in-focus image completion approach is presented in the Fig. 10. Feeding multiple images at different synthetic focal planes, it can be observed that our method can output an all-in-focus image from blurring images. Compared with the raw input image, the generated hyper-spectral image is more clear in the boundary and details.

5 Conclusion

This paper proposes the first unsupervised learning framework to train the depth-from-defocus and depth-from-focus neural networks simultaneously to estimate scene depth. The framework learning process is guided by the depth consistency between depth-from-defocus and depth-from-focus, as well as the defocus consistency between the recovered defocus image and the original defocus image input. In real applications depth-from-defocus and depth-from-focus can separately estimate the depth map based on a single image or image stack, which overcomes the scale issue commonly existed in monocular camera depth estimation.

References

Sameer Agarwal, Noah Snavely, Ian Simon, Steven M Seitz, and Richard Szeliski. Building rome in a day. In *IEEE International Conference on Computer Vision*, pp. 72–79, 2009.

Sameer Agarwal, Yasutaka Furukawa, Noah Snavely, Ian Simon, Brian Curless, Steven M Seitz, and Richard Szeliski. Building rome in a day. *Communications of the ACM*, 54(10):105–112, 2011.

Ibraheem Alhashim and Peter Wonka. High quality monocular depth estimation via transfer learning. *arXiv preprint arXiv:1812.11941*, 2018.

V Madhu Babu, Kaushik Das, Anima Majumdar, and Swagat Kumar. Undemon: Unsupervised deep network for depth and ego-motion estimation. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1082–1088. IEEE, 2018.

Yuanzhouhan Cao, Zifeng Wu, and Chunhua Shen. Estimating depth from monocular images as classification using deep fully convolutional residual networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(11):3174–3182, 2017.

Marcela Carvalho, Bertrand Le Saux, Pauline Trouvé-Peloux, Andrés Almansa, and Frédéric Champagnat. Deep depth from defocus: how can defocus blur improve 3d estimation using dense neural networks? In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 0–0, 2018.

David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE international conference on computer vision*, pp. 2650–2658, 2015.

David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *NIPS*, 2014.

Ravi Garg, Vijay Kumar BG, Gustavo Carneiro, and Ian Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *ECCV*, pp. 740–756, 2016.

Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, 2017.

Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel Brostow. Digging into self-supervised monocular depth estimation. *ICCV*, 2019.

Qi Guo, Zhujun Shi, Yao-Wei Huang, Emma Alexander, Cheng-Wei Qiu, Federico Capasso, and Todd Zickler. Compact single-shot metalens depth sensors inspired by eyes of jumping spiders. *Proceedings of the National Academy of Sciences*, 116(46):22959–22965, 2019.

Shir Gur and Lior Wolf. Single image depth estimation trained via depth from defocus cues. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7683–7692, 2019.

Samuel W Hasinoff and Kiriakos N Kutulakos. A layer-based restoration framework for variable-aperture photography. In *2007 IEEE 11th International Conference on Computer Vision*, pp. 1–8. IEEE, 2007.

Caner Hazirbas, Sebastian Georg Soyer, Maximilian Christian Staab, Laura Leal-Taixé, and Daniel Cremers. Deep depth from focus. In *Asian Conference on Computer Vision*, pp. 525–541. Springer, 2018.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

Hyungjoo Jung, Youngjung Kim, Dongbo Min, Changjae Oh, and Kwanghoon Sohn. Depth prediction from a single image with conditional adversarial networks. In *2017 IEEE International Conference on Image Processing (ICIP)*, pp. 1717–1721. IEEE, 2017.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105, 2012.

Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *3DV*, pp. 239–248, 2016.

Jae-Han Lee, Minhyeok Heo, Kyung-Rae Kim, and Chang-Su Kim. Single-image depth estimation based on fourier domain analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 330–339, 2018.

Haiting Lin, Can Chen, Sing Bing Kang, and Jingyi Yu. Depth recovery from light field using focal stack symmetry. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3451–3459, 2015.

Fayao Liu, Chunhua Shen, and Guosheng Lin. Deep convolutional neural fields for depth estimation from a single image. In *CVPR*, 2015.

Shaojun Liu, Fei Zhou, and Qingmin Liao. Defocus map estimation from a single image based on two-parameter defocus model. *IEEE Transactions on Image Processing*, 25(12):5943–5956, 2016.

Michael Moeller, Martin Benning, Carola Schönlieb, and Daniel Cremers. Variational depth from focus reconstruction. *IEEE Transactions on Image Processing*, 24(12):5369–5378, 2015.

Takashi Nagata, Mitsumasa Koyanagi, Hisao Tsukamoto, Shinjiro Saeki, Kunio Isono, Yoshinori Shichida, Fumio Tokunaga, Michiyo Kinoshita, Kentaro Arikawa, and Akihisa Terakita. Depth perception from image defocus in a jumping spider. *Science*, 335(6067):469–471, 2012.

Matteo Poggi, Fabio Tosi, and Stefano Mattoccia. Learning monocular depth estimation with unsupervised trinocular assumptions. In *2018 International Conference on 3D Vision (3DV)*, pp. 324–333. IEEE, 2018.

Rene Ranftl, Vibhav Vineet, Qifeng Chen, and Vladlen Koltun. Dense monocular depth estimation in complex dynamic scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4058–4066, 2016.

Sean Ryan Fanello, Christoph Rhemann, Vladimir Tankovich, Adarsh Kowdle, Sergio Orts Escolano, David Kim, and Shahram Izadi. Hyperdepth: Learning depth from structured light without matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5441–5450, 2016.

Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European Conference on Computer Vision*, pp. 746–760. Springer, 2012.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Pratul P Srinivasan, Tongzhou Wang, Ashwin Sreelal, Ravi Ramamoorthi, and Ren Ng. Learning to synthesize a 4d rgbd light field from a single image. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2243–2251, 2017.

Deqing Sun, Stefan Roth, and Michael J Black. Secrets of optical flow estimation and their principles. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pp. 2432–2439. IEEE, 2010.

Supasorn Suwajanakorn, Carlos Hernandez, and Steven M Seitz. Depth from focus with your mobile phone. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3497–3506, 2015.

Chris Sweeney, Torsten Sattler, Tobias Hollerer, Matthew Turk, and Marc Pollefeys. Optimizing the viewing graph for structure-from-motion. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 801–809, 2015.

Benjamin Ummenhofer, Huizhong Zhou, Jonas Uhrig, Nikolaus Mayer, Eddy Ilg, Alexey Dosovitskiy, and Thomas Brox. Demon: Depth and motion network for learning monocular stereo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5038–5047, 2017.

Diana Wofk, Fangchang Ma, Tien-Ju Yang, Sertac Karaman, and Vivienne Sze. Fastdepth: Fast monocular depth estimation on embedded systems. In *2019 International Conference on Robotics and Automation (ICRA)*, pp. 6101–6108. IEEE, 2019.

Dan Xu, Elisa Ricci, Wanli Ouyang, Xiaogang Wang, and Nicu Sebe. Multi-scale continuous crfs as sequential deep networks for monocular depth estimation. In *CVPR*, 2017.

Zhichao Yin and Jianping Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1983–1992, 2018.

Hang Zhao, Orazio Gallo, Iuri Frosio, and Jan Kautz. Loss functions for neural networks for image processing. *arXiv preprint arXiv:1511.08861*, 2015.

Changyin Zhou, Stephen Lin, and Shree Nayar. Coded aperture pairs for depth from defocus. In *2009 IEEE 12th international conference on computer vision*, pp. 325–332. IEEE, 2009.

Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1851–1858, 2017.

Shaojie Zhuo and Terence Sim. Defocus map estimation from a single image. *Pattern Recognition*, 44(9):1852–1858, 2011.

Yuliang Zou, Zelun Luo, and Jia-Bin Huang. Df-net: Unsupervised joint learning of depth and flow using cross-task consistency. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 36–53, 2018.