Deep Depth from Defocus: How Can Defocus Blur Improve 3D Estimation Using Dense Neural Networks?

Marcela Carvalho[1](✉), Bertrand Le Saux[1], Pauline Trouvé-Peloux[1], Andrés Almansa[2], and Frédéric Champagnat[2]

[1] DTIS, ONERA, Université Paris-Saclay, 91123 Palaiseau, France
{marcela.carvalho,bertrand.saux,pauline.trouve}@onera.fr
[2] Université Paris Descartes, 75006 Paris, France
{andres.almansa,frederic.champagnat}@parisdescartes.fr

Abstract. Depth estimation is critical interest for scene understanding and accurate 3D reconstruction. Most recent approaches with deep learning exploit geometrical structures of standard sharp images to predict depth maps. However, cameras can also produce images with defocus blur depending on the depth of the objects and camera settings. Hence, these features may represent an important hint for learning to predict depth. In this paper, we propose a full system for single-image depth prediction in the wild using depth-from-defocus and neural networks. We carry out thorough experiments real and simulated defocused images using a realistic model of blur variation with respect to depth. We also investigate the influence of blur on depth prediction observing model uncertainty with a Bayesian neural network approach. From these studies, we show that out-of-focus blur greatly improves the depth-prediction network performances. Furthermore, we transfer the ability learned on a synthetic, indoor dataset to real, indoor and outdoor images. For this purpose, we present a new dataset with real all-focus and defocused images from a DSLR camera, paired with ground truth depth maps obtained with an active 3D sensor for indoor scenes. The proposed approach is successfully validated on both this new dataset and standard ones as NYUv2 or Depth-in-the-Wild. Code and new datasets are available at https://github.com/marcelampc/d3net_depth_estimation.

Keywords: Depth from defocus · Domain adaptation · Depth estimation · Single-image depth prediction

1 Introduction

3D reconstruction has a large field of applications such as in human computer interaction, augmented reality and robotics, which have driven research on the topic. This reconstruction usually relies on accurate depth estimates to process the 3D shape of an object or a scene. Traditional depth estimation approaches

L. Leal-Taixé and S. Roth (Eds.): ECCV 2018 Workshops, LNCS 11129, pp. 307–323, 2019.
https://doi.org/10.1007/978-3-030-11009-3_18

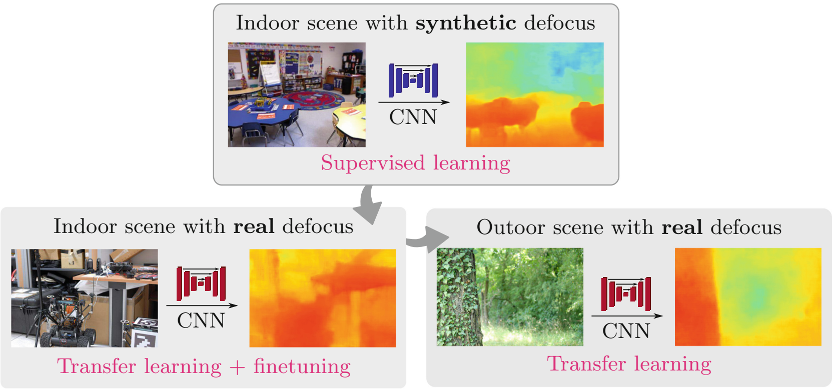

Indoor scene with **synthetic** defocus
CNN
Supervised learning
Indoor scene with **real** defocus
CNN
Transfer learning + finetuning
Outoor scene with **real** defocus
CNN
Transfer learning

Fig. 1. Depth estimation with synthetic and real defocused data on indoor and outdoor challenging scenes. These results show the flexibility to new datasets of a model trained with a synthetically defocused indoor dataset, finetuned on a real DSLR indoor set and finally tested in outdoor scenes without further training.

exploit different physical aspects to extract 3D information from perception, such as stereoscopic vision, structure from motion, structured light and other depth cues in 2D images [1,2]. However, some of these techniques depend on the environment (*e.g.* sun, texture) or even require several devices (*e.g.* camera, projector), leading to cumbersome systems. Many efforts have been made to make them compact: *e.g.* the light-field cameras which use a microlens array in front of the sensor, from which a depth map can be extracted [3] (Fig. 1).

In recent years, several approaches for depth estimation based on deep learning (deep depth estimation), have been proposed [4]. These methods use a single image and thus lead to compact, standard systems. Most of them exploit depth cues in the image based on geometrical aspects of the scene to estimate the 3D structure with the use of convolutional neural networks (CNNs) [5–8]. A few ones can also make use of additional depth cues such as stereo information to train the network [9] and improve predictions.

Another important cue for depth estimation is defocus blur. Indeed, Depth from Defocus (DFD) has been widely investigated [10–15]. It led to various analytical methods and corresponding optical systems for depth prediction. However, conventional DFD suffers from ambiguity in depth estimation with respect to the focal plane and dead zone, due to the camera depth of field where no blur can be measured. Moreover, DFD requires a scene model and an explicit calibration between blur level and depth value to estimate 3D information. Thus, it is tempting to integrate defocus blur with the power of neural networks, which leads to the question: does defocus blur improve deep depth estimation performances?

In this paper, we use a dense neural network, D3-Net [16], in order to study the influence of defocus blur on depth estimation. First it is tested on a synthetically defocused dataset created from NYUv2 with optically realistic blur variation, which allows to compare several optical settings. We further examine the uncertainty of the CNN predictions with and without blur. We then explore real defocused data with a new dataset which comprises indoor all-in-focus and defocused images, and corresponding depth maps. Finally, we verify how the deep model behaves when confronted to challenging images in the wild with the Depth-in-the-Wild [17] dataset and further outdoor images.

These experiments show that defocused information is exploited by neural networks and is indeed an important hint to improve deep depth estimation. Moreover, the joint use of structural and blur information proposed in this paper overcomes current limitations of single-image DFD. Finally, we show that these findings can be used in a dedicated device with real defocus blur to actually predict depth indoors and outdoors with good generalization.

2 Related Work

Deep Monocular Depth Estimation. Several works have been developed to perform monocular depth estimation based on techniques of machine learning. One of the first solutions was presented by Saxena *et al.* [18], which formulate the depth estimation for the Make3D dataset as a Markov Random Field (MRF) problem with horizontally aligned images using a multi-scale architecture. More recent solutions are based on CNNs to exploit spatial correlation by enforcing a local connectivity. Eigen *et al.* [4,5] proposed a multi-scale architecture capable of extracting global and local information from the scene. In [19], Cao *et al.* used a Conditional Random Field (CRF) to post-process the output of a deep residual network (ResNet) [20] in order to improve the reliability of the predictions. Xu *et al.* [21] adopted a deeply supervised approach connecting intermediate outputs of a ResNet to a continuous CRF fusion module to combine depth prediction at different scales achieving higher performance. Also adopting residual connections, Laina *et al.* [22] proposed an encoder-decoder architecture with fast up-projection blocks. More recently, Jung *et al.* [23] introduced generative adversarial networks [24] (GANs) adapting an adversarial loss to refine the depth map predictions. With a different strategy, [9,25,26] propose to investigate the epipolar geometry using CNNs. DeMoN [9] jointly estimates a depth map and camera motion given a sequential pair of images with optical flow. [25,26] use unsupervised learning to reconstruct stereo information and predict depth. More recently, Kendall and Gal [27] and Carvalho *et al.* [16] explore the reuse of feature maps, building upon an encoder decoder with dense and skip connections [28]. While [27] propose a regression function that captures the uncertainty of the data, [16] uses an adversarial loss.

The aforementioned techniques for monocular depth estimation with neural networks base their learning capabilities on structured information (*e.g.*, textures, linear perspective, statistics of objects and their positions). However, depth perception can use another well-know cue: defocus blur. We present in the following section state-of-the-art approaches from this domain.

Depth Estimation Using DFD. In computational photography, several works investigated the use of defocus blur to infer depth [10]. Indeed, the amount of defocus blur of an object can be related to its depth using geometrical optics $\epsilon = Ds \cdot \left| \frac{1}{f} - \frac{1}{d_{out}} - \frac{1}{s} \right|$, where f stands for the focal length, d_{out} the distance of the object with respect to the lens, s the distance between the sensor and the lens and D the lens diameter. $D = {}^{f}/{}_{N}$, where N is the f-number (Fig. 2).

Recent works usually use DFD with a single image (SIDFD). Although the acquisition is simple, it leads to more complex processing as both the scene and the blur are unknown. State of the art approaches use analytical models for the scene such as sharp edges models [15] or statistical scene Gaussian priors [12,29]. Coded apertures have also been proposed to improve depth estimation accuracy [11,14,30,31].

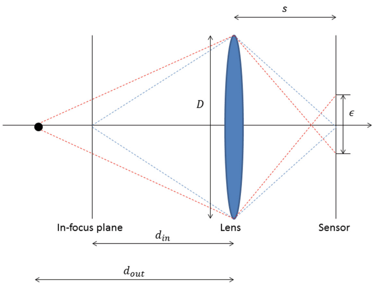


Fig. 2. Illustration of the DFD principle. Rays originating from the out of focus point (black dot) converge before the sensor and spread over a disc of diameter ϵ.

Nevertheless, SIDFD suffers from two main limitations: first, there is an *ambiguity* related to the object's position in front or behind the in-focus plane; second, blur variation cannot be measured in the camera depth of field, leading to a *dead zone.* Ambiguity can be solved using asymmetrical coded aperture [14], or even by setting the focus at infinity. Second, dead zones can be overcome using several images with various in-focus planes. In a single snapshot context, this can be obtained with unconventional optics such as a plenoptic camera [32] or a lens with chromatic aberration [12,33], but both at the cost of image quality (low resolution or chromatic aberration).

Indeed, inferring depth from the amount of defocus blur with model-based techniques requires a tedious explicit calibration step, usually conducted using point sources or a known high frequency pattern [11,34] at each potential depth. These constraints lead us to investigate data-based methods using deep learning techniques to explore structured information together with blur cues.

Learning Depth from Defocus Blur. The existence of common datasets for depth estimation [1,32,35], containing pairs of RGB images and corresponding depth maps, facilitates the creation of synthetic defocused images using real camera parameters. Hence, a deep learning approach can be used. To the best of our knowledge, only a few papers in the literature use defocus blur as a cue in learning depth from a single image. Srinivasan *et al.* [36] uses defocus blur to train a network dedicated to monocular depth estimation: the model measures the consistency of simulated defocused images, generated from the estimated depth map and all-in-focus image, with true defocused images. However, the final network is used to conduct depth estimation from all-in-focus images. Hazirbas *et al.* [32] propose to use a focal stack, which is more related to depth from focus approaches than DFD. Finally, [37] presents a network for depth estimation and deblurring using a single defocused image. This work shows that networks can integrate blur interpretation. However, [37] creates a synthetically defocused dataset from real NYUv2 images without a realistic blur variation with respect to the depth, nor sensor settings (*e.g.*, camera aperture, focal distance). However, there has not been much investigation about how defocus blur influence on depth estimation, nor how can these experiments improve depth prediction in the wild.

In contrast to previous works, to the best of our knowledge, we present the first system for deep depth from defocus (Deep-DFD): *i.e.* single-image depth prediction in the wild using deep learning and depth-from-defocus. In Sect. 3, we study the influence of defocus blur on deep depth estimation performances. (i) We run tests on a synthetically defocused dataset generated from a set of true depth maps and all-in-focus images. The amount of defocus blur with respect to depth varies according to a physical optical model to relate to realistic examples. (ii) We also compare performances of deep depth estimation with several optical settings: we compare all-in-focus images with defocused images of three different focus settings. (iii) We analyse the influence of defocus blur on neural networks using uncertainty maps and diagrams of errors per depth. In Sect. 4, (iv) we carry out validation and analysis on a new dataset created with a Digital Single Lens Reflex (DSLR) camera and a calibrated RGB-D sensor. At last, in Sect. 5, (v) we show the network is able to generalize to images in the wild.

3 Learning DFD to Improve Depth Estimation

In this section, we perform a series of experiments with synthetic and real defocused data exploring the power of deep learning to depth prediction. As we are interested in using blur as a cue, we do not apply any image processing for data augmentation capable of modifying out-of-focus information. Hence, for all experiments, we extract random crops of 224 × 224 from the original images and apply horizontal flip with a probability of 50%. Tests are realized using the full-resolution image.

3.1 D3-Net Architecture

To perform such tests, we adopt the D3-Net architecture from [16], illustrated in Fig. 3. We use the PyTorch framework on a NVIDIA TITAN X GPU with 12 GB of memory. We initialize the D3-Net encoder, corresponding to DenseNet-121, with pretrained weights on Imagenet dataset and D3-Net decoder with random weights from a normal distribution with zero-mean and 0.2 variance. We add dropout [39] regularization with a probability of 0.5 to the first four convolutional layers of the decoder as we noticed it improves generalization. We also adopt dropout layers to posteriorly study model's uncertainty.

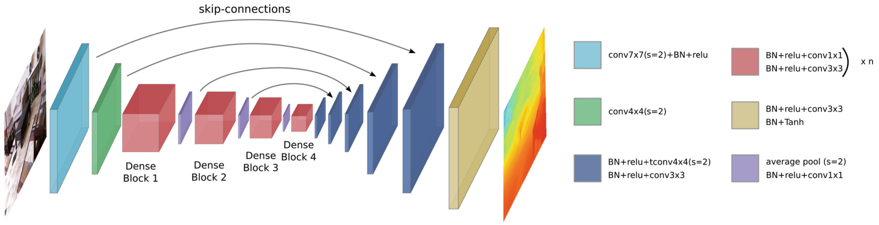


Fig. 3. D3-Net architecture from [16]. The encoder part corresponds to a DenseNet-121 [28], with $n = 6, 12, 24, 16$, respectively for indicated Dense Blocks. The encoder-decoder structure is based on U-Net [38] to explore the reuse of feature maps.

3.2 Synthetic NYUv2 with Defocus Blur

The NYU-Depth V2 (NYUv2) dataset [35] has approximately 230k pairs of images from 249 scenes for training and 215 scenes for testing. In [16], D3-Net reaches its best performances when trained with the complete dataset. However, NYUv2 also contains a smaller split with 1449 pairs of aligned RGB and depth images, of which 795 pairs are used for training and 654 pairs for testing. Therefore, experiments in this section were performed using this smaller dataset to fasten experiments. Original frames from Microsoft Kinect output have the resolution of 640 × 480. Pairs of images from the RGB and Depth sensors are posteriorly aligned, cropped and processed to fill-in invalid depth values. Final resolution is 561 × 427.

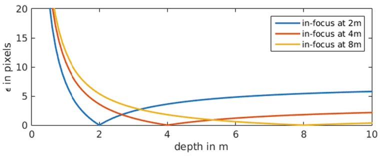


Fig. 4. Blur diameter variation vs depth for the in-focus settings: 2 m, 4 m and 8 m tests on the NYUv2.

To generate physically realistic out-of-focus images, we choose the parameters corresponding to a synthetic camera with a focal length of 15 mm, f-number 2.8 and pixel size of 5.6 µm. Three settings of in-focus plane are tested, respectively at 2 m, 4 m and 8 m from the camera. Figure 4 shows the variation of the blur diameter ϵ with respect to depth, for both settings and Fig. 5 shows examples of synthetic defocused images. As illustrated in Fig. 4, setting the in-focus plane at 2 m corresponds to a camera with small depth of field. The objects in the depth range from 1 to 10 m will present small defocus blur amounts, apart from the objects in the camera depth of field, which remain sharp. Note that this configuration suffers from depth ambiguity caused by the blur estimation. Setting the in-focus plane at a larger depth, here 4 m or 8 m, corresponds to a camera with larger depth of field. Only the closest objects will show defocus blur, with a comparatively larger blur ammount between 0–3 m than previous setting. This can be observed in the extracted details in Fig. 5.

To create the out-of-focus dataset, we adopt the layered approach of [40] where each defocused image $\widehat{L}$ is the sum of K blurred images multiplied by masks, A_k, related to local object depth, k, and occlusion of foreground objects:

$$\widehat{L} = \sum_k \left[(A_k L + A_k^* L_k^*) * h(k)) \right] M_k, \tag{1}$$

where $h(k)$ is the defocus blur at distance k, L is the all-in-focus image and $A_k^* L_k^*$, the layer extension behind occluders, obtained by inpainting. Finally M_k models the cumulative occlusions defined as:

$$M_k = \prod_{k'=k+1}^{K} (1 - A_{k'} * h(k')). \tag{2}$$

Following [36], we chose to model the blur as a disk function which the diameter varies with the depth.

As will be discussed later, the proposed approach can be disputable as the true depth map is used to generate the out-of focus image. However, this strategy allows us easily perform various experiments to analyze the influence of blur corresponding to different in-focus settings in the image.

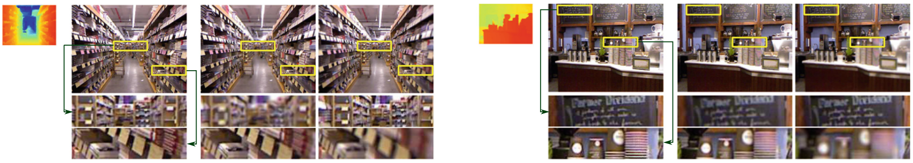


Fig. 5. Examples of synthetic defocused images generated from an image of the NYUv2 database for two camera in-focus plane settings: 2 and 8 m.

3.3 Performance Results

Table 1 shows performance of D3-Net first using all-in-focus and then defocused images with proposed settings. As illustrated in Fig. 4, when the in-focus plane is at 8 m, there is no observable ambiguity. Hence performance comparison with SIDFD methods can then be made. So, we include the performances of two methods from the SIDFD literature [15,41] which estimate the amount of local blur using either sharp edge model or gaussian prior on the scene gradients.

Several conclusions can be drawn from Table 1. First, as already stated by Anwar *et al.*, there is a significant improvement on depth estimation when using out-of-focus images instead of all-in-focus images. Second, D3-Net outperforms the standard model-based SIDFD methods, which can also be observed in Fig. 8, without requiring an analytical scene model nor explicit blur calibration. Furthermore, there is also a sensitivity of the depth estimation performance with respect to the position of the in-focus plane. The best setting for these tests is

Table 1. Performance comparison of D3-Net using all-in-focus and defocused images with different settings, and two SIDFD approaches [15,41] for the 8 m focus setting.

Methods	Error↓				Accuracy↑		
	rel	log10	rms	rmslog	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Original RGB images							
D3-Net All-in-focus	0.226	-	0.706	-	65.8%	89.2%	96.7%
RGB images with additional blur							
D3-Net 2 m focus	0.068	0.028	0.274	0.110	96.1%	99.0%	99.6%
D3-Net 4 m focus	0.085	0.036	0.398	0.125	92.5%	99.0%	99.8%
D3-Net 8 m focus	0.060	-	0.324	-	95.2%	99.1%	99.9%
Zhuo *et al.* [15] 8 m focus	0.273	-	0.981	-	51.7%	83.1%	95.1%
Trouvé *et al.* [41] 8 m focus	0.429	0.289	1.743	0.956	39.2%	52.7%	61.5%
RGB images with additional blur proposed by [37]							
Anwar *et al.* [37]	0.094	0.039	0.347	-	-	-	-
D3-Net	0.036	0.016	0.144	0.054	99.3%	100.0%	100.0%

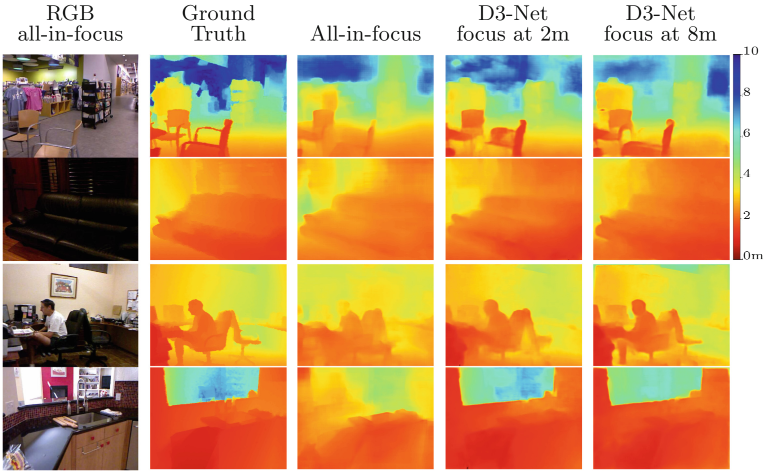


Fig. 6. Qualitative comparison for different predictions with the proposed defocus blur configurations.

with the in-focus plane at 2 m. This corresponds to a significant amount of blur for most of the objects but near the focal plane. And shows that the network actually uses blur cues and is able to overcome depth ambiguity using geometrical structural information. Figure 8 also illustrates this conclusion: the scene has mainly three depth levels with a foreground, a background, and an intermediate level around 2 m. The corresponding out-of-focus image is generated using an in-focus plane at 2 m. Using [15], the background and the foreground are at the same depth, while D3-Net shows no such error in the depth map.

Finally, we also trained and tested D3-Net with the dataset proposed in [37]. However, differently from the method explored in our paper, the out-of-focus images were generate without any regard to camera settings. The last two lines from Table 1 shows that D3-Net also outperforms the network in [37].

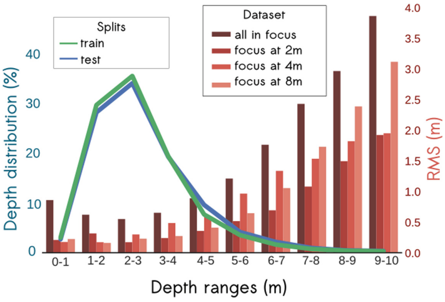
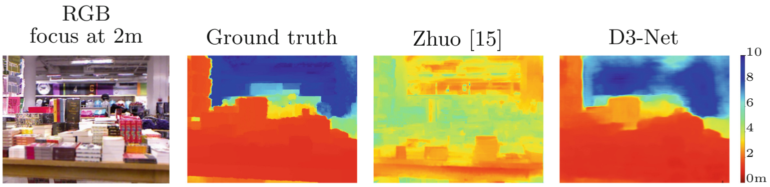


Fig. 7. Distribution of pixels on different depth ranges and RMS performance of D3-Net trained with and without defocus blur.

Also, Fig. 6 and columns 3 and 6 from Fig. 9 show that estimations from out-of-focus images are sharper than from all-in-focus images. Indeed, defocus blur provides extra local information to the network leading to a better depth segmentation.

Per Depth Error Analysis. There is an intrinsic relation between the number of examples a network can learn from and its performance when tested on similar samples. Here, we compare the prediction error per depth range between all-in-focus and defocused images. We observe the relation to depth data distribution. Figure 7 shows in the same plot repartition the RMS per depth in meters and the depth distribution for testing and training images with NYUv2.

For all-in-focus images, the errors seem to be highly correlated to the number of examples in the dataset. Indeed, a minimum error is obtained for 2 m, corresponding to the depth with the highest number of examples. On the other hand, using defocus blur, errors repartition is more similar to a quadratic increase of error with depth, which is the usual error repartition of passive depth estimation.


RGB focus at 2m

Ground truth

Zhuo [15]

D3-Net

Fig. 8. Comparison between D3-Net estimation and Zhuo [15] for images with the focus plane at 2 m.

Furthermore, the 2 m focus setting does not show an error increase at its focal plane position, though it corresponds to the dead zone of SIDFD. This surprising result shows that the proposed approach overcomes this issue probably because the neural network also relies on context and geometric features. In general, 2 m, 4 m and 8 m focus have similar performance for depth range between 0 to 3 m. After this depth, the 2 m focus presents the lowest errors. When focus is at 4 m,

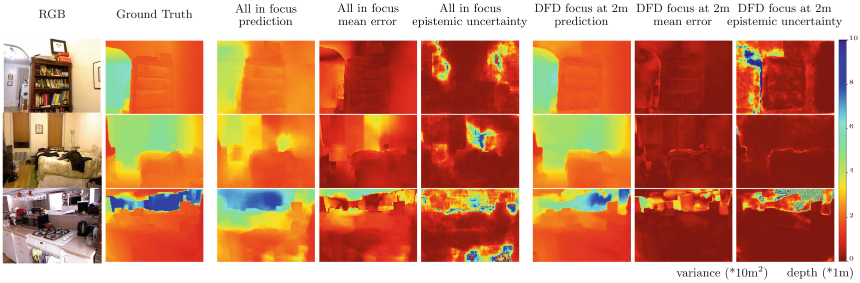


Fig. 9. Qualitative comparison of all focus and DFD with 2 m focus prediction, mean error and epistemic uncertainty with NYUv2 dataset. Lower values of depth and uncertainties are represented by warmer colors. (Color figure online)

we observe a drop in all metrics performances compared to 2 m and 8 m. The reason can be observed when comparing both Figs. 4 and 7. This configuration presents worst RMS performances between 3 and 7 m, when blur information is too small to be used by the network and there is not enough data to overcome the missing cue, but enough to worsen results. The same happens to the model at 8 m, where results are more prone to errors after approximately 7 m.

3.4 Uncertainties on the Depth Estimation Model

To go further in the analysis of understanding the influence of blur in depth prediction, we present a study on model uncertainties following [27,42,43]. More precisely, we evaluate the epistemic uncertainty of the deep network model, or how ignorant is the model with respect to the dataset probabilistic distribution.

To perform this experiment, we place a prior distribution over the network weights to replace the deterministic weight parameters at test time [27]. We adopt the Monte Carlo dropout method [43] to measure variational inference placing dropout layers during train and also during test phases. Following [42], we produce 50 samples for each image, calculate the mean prediction and the variance of these predictions to generate the model uncertainty.

Figure 9 presents examples of the network prediction, mean error and epistemic uncertainty for the NYUv2 dataset with sharp images and with focus at 2 m. Mean error is produced using the ground truth image, while the variance only depends on the model's prior distribution. For both configurations, highest variances are observed in non-textured areas and edges, as predictable. However, the model with blur has less diffuse uncertainty: it is concentrated on the object edges, and these objects are better segmented. In the second row of the figure, we observe that the all-in-focus model has difficulties to find an object near the window, while this is overcome with blur cues present on the defocused model. In the first row, we observe high levels of uncertainty at the zones near the bookcase, defocused model reduce some of this variance with defocus information. Finally, the last row presents a hard example where both models have high prediction variances mainly in the top middle part, where there is a hole. However the all-in-focus model also presents high mean error and variance in the bottom zone unlike the model with blur.

4 Experiments on a Real Defocused Dataset

In Sect. 3, several experiments were performed using a synthetic version of NYUv2. However, when adopting convolutional neural networks, it can be a little tricky to use the desired output (depth) to create blur information on the input of the network. So, in this section, we propose to validate our method on real defocused data from a DSLR camera paired with the respective depth map from a calibrated RGB-D sensor.

Dataset Creation. To create a DFD dataset, we paired a DSLR Nikon D200 with an Asus Xtion sensor to produce out-of-focus data and corresponding depth maps, respectively. Our platform can be observed in Fig. 10. We carefully calibrate the depth sensor to the DSLR coordinates to produce RGB images alligned with the corresponding depth map. The proposed dataset contains 110 images from indoor scenes, with 81 images for training and 29 images for testing. Each scene is acquired with two camera apertures: $N = 2.8$ and $N = 8$, providing respectively out-of-focus and all-in-focus images.

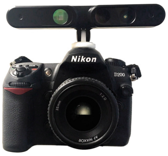


Fig. 10. Experimental platform with Xtion PRO sensor coupled to a DSLR Nikon camera.

As the DFD dataset contains a small amount of images, we pretrain the network using simulated images from NYUv2 dataset and then conduct a finetuning of the network using the real dataset. The DSLR camera originally captures images of high resolution 3872 × 2592; but to reduce the calculation burden, we downsample them to 645 × 432. In order to simulate defocused images from NYUv2 as similar as possible to DSLR's, the images from the Kinect are upsampled and cropped to have the same resolution and the same field of view as the downsampled DSLR images. Then defocus blur is applied to the images using the same method as in Sect. 3 but with a blur variation with that fits the real blur variation of the DSLR, obtained experimentally.

Performance Results. Using the new dataset, we perform three experiments: first we train D3-Net with the in-focus and defocused dataset respectively, using same patch approach from last experiments. We also test D3-Net with the in-focus dataset using an strategy that explores the global information of the scene and a series of preprocessing methods: we resize input images to 320 × 256 and performance data augmentation suggested in [4] to improve generalization.

In Table 2, the performances from the proposed models can be compared. The results show that defocus blur does improve the network performance increasing 10 to 20 percentual points in accuracy and also gives qualitative results with better segmentation as illustrated in Fig. 11.

The network is capable to find a relation between depth and defocus blur and predict better results, even thought the it may miss from global information

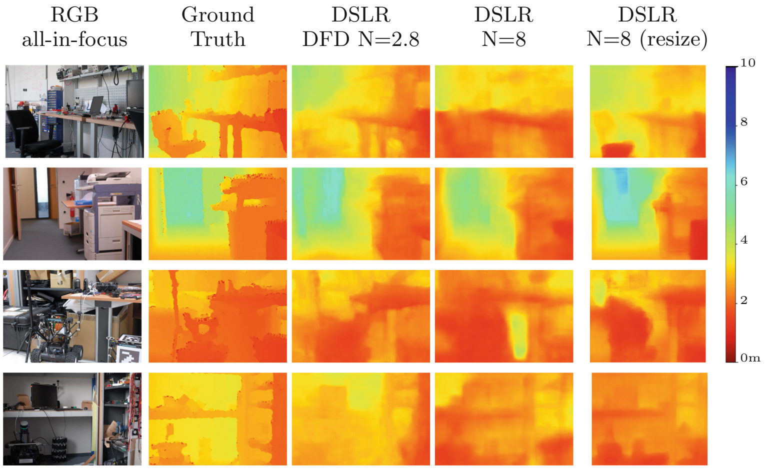


Fig. 11. Qualitative comparison of D3-Net trained on defocused and all-focused images from a DSLR camera.

Table 2. Performance comparison of D3-Net using all-in-focus and defocused images on a real DSLR dataset.

Methods	Error↓				Accuracy↑		
	rel	log10	rms	rmslog	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
$N = 2.8$	0.157	0.065	0.546	0.234	80.9%	94.4%	97.6%
$N = 8$	0.225	0.095	0.730	0.285	60.2%	87.7%	98.0%
$N = 8$ (resize)	0.199	0.084	0.654	0.259	69.6%	91.6%	97.4%

when being trained with small patches. When feeding the network with resized images, filters from the last layers of the encoder, as from the first layers of the decoder, can understand the global information as they are fed with feature maps from the entire scene in a low resolution. However, this relation is not enough to give better predictions. As we can observe in the first examples of the third row in Fig. 11, the DFD D3-Net used defocus to find the contours of the object, meanwhile the standard D3-Net wrongly predicts the form of a chair, as it is an object constantly present in front of a desk. Our experiments show that the Deep-DFD model is more robust to generalization and less prone to overfitting than traditional methods trained and finetuned on all-in-focus images.

5 Depth “in the Wild”

In the era of autonomous driving vehicles (on land, on water, or in the air), there has been an increasing demand of less intrusive, more robust sensors and processing techniques to embed in systems able to evolve in the wild. Previously, we validated our approach with several experiments on indoor scenes and we proved that blur can be learned by a neural network to improve prediction and also to improve the model's confidence to its estimations. In this section, we now propose to tackle the general case of uncontrolled scenes. We first assess the ability of the standard D3-Net, trained without defocus blur, to generalize to "in-the-wild" images using the Depth-in-the-Wild dataset [17] (DiW). Second, we use the whole system, D3-Net trained on indoor defocused images and the DSLR camera described from Sect. 4, in uncontrolled, outdoor environments.

Depth-in-the-Wild Dataset (DiW). The ground truth of the DiW dataset is not dense; indeed, only two points of each RGB image are relatively annotated as being closer or farther from the camera, or at the same distance. To adapt the network, we replace the objective function of D3-Net by the one proposed by the authors of the dataset [17]. Then, for training, we take the weights of D3-Net trained on all-in-focus NYUv2 [16], and finetune the model on DiW using the modified network. We show the results of this model on the test set of DiW in Fig. 12. The predicted depths present sharp edges for people and objects and give plausible estimates of the 3D structure of the given scenes. However, as the network was mostly trained on indoor scenes, it cannot give accurate depth predictions on sky regions. This shows that the a neural network has inherent capacity to predict depth in the wild. We will now see that we can improve this capacity by integrating physical cues of the sensor.

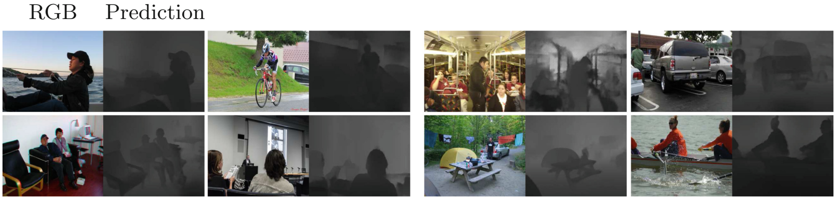


Fig. 12. Examples of depth prediction using DIW dataset with D3-Net trained on NYUv2.

Deep-DFD in the Wild. We now observe how deep models trained with blurred indoor images behave when confronted to challenging outdoor scenes. These experiments explore the model's capability to adapt predictions to new scenarios, never seen during training. To perform our tests, we first acquire new data using the DSLR camera with defocus optics (from Sect. 4) and keeping the same camera settings. As the depth sensor from the proposed platform works poorly outdoor, this new set of images does not contain respective depth ground truth. Thus, the model is neither trained on the new data, nor finetuned. Indeed, we use directly the models finetuned on indoor data with defocus blur (Sect. 4).

Results from the CNN models and from Zhuo's [15] analytical method are shown in Fig. 13. With D3-Net trained on all-in-focus images, the model constantly fails to extract information from new objects, as can be observed in the images with the road and also with the tree trunk. As expected, this model tries to base prediction on objects similar to what those seen during training or during finetuning, which are mostly non-existent in these new scenes. On the contrary, though the model trained with defocus blur information has equally never seen these new scenarios, the predictions give results relatively close to the expected depth maps. Indeed, the Deep-DFD model notably extracts and uses blur information to help prediction, as geometric features are unknown for the trained network. Finally, Zhuo's method also gives encouraging results, but constantly fails duo to defocus blur ambiguity to the focal plane (as on the handrail on the top left example of Fig. 13). As can be deduced from our experiments, the combined use of geometric, statistical and defocus blur is a promising method to generalize learning capabilities.

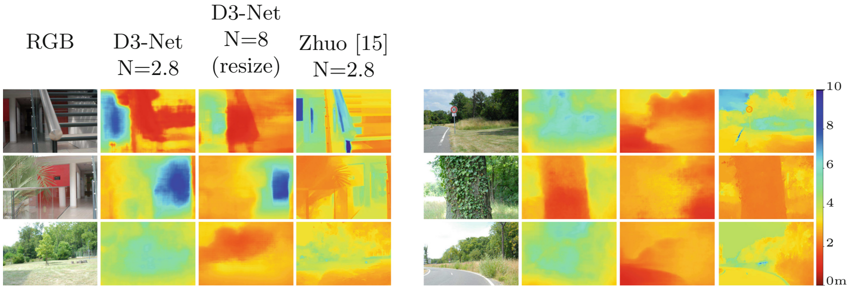


Fig. 13. Depth estimation methods: from left to right, D3-Net trained on defocused images, all-in-focus images and a classical Depth from Defocus approach by [15].

6 Conclusion

In this paper, we have studied the influence of defocus blur as a cue in a monocular depth estimation using a deep learning approach. We have shown that the use of blurred images outperforms the use of all-in-focus images, without requiring any scene model nor blur calibration. Besides, the combined use of defocus blur and geometrical structure information on the image, brought by the use of a deep network, avoids the classical limitations of DFD with a conventional camera (*e.g.*, depth ambiguity, dead zones). We have proposed different tools to visualize the benefit of defocus blur on the network performance, such as per depth error statistics and uncertainty maps. These tools have shown that depth estimation with defocus blur is most significantly improved at short depths, resulting in better depth map segmentations. We have also compared performance of Deep-DFD with several optical settings to better understand the influence of the camera parameters to deep depth prediction. In our tests, the best performances were obtained for a close in-focus plane, which leads to really small camera depths of field and thus defocus blur on most of the objects in the dataset.

Besides synthetic data, this paper also provides excellent results on both indoor and outdoor real defocused images from a new set of DSLR images. These experiments on real defocused data proved that defocus blur combined to neural networks are more robust to training data and domain generalization, reducing possible constraints of actual acquisition models with active sensors and stereo systems. Notably, results on the challenging domain of outdoor scenes without further calibration, or finetuning prove that this new system can be used in the wild to combine physical information (defocus blur) and geometry and perspective cues already used by standard neural networks. These observations open the way to further studies on the optimization of the camera parameters and acquisition modalities for 3D estimation using defocus blur and deep learning.

References

1. Saxena, A., Sun, M., Ng, A.Y.: Make3D: learning 3D scene structure from a single still image. IEEE Trans. Pattern Anal. Mach. Intell. **31**(5), 824–840 (2009)
2. Calderero, F., Caselles, V.: Recovering relative depth from low-level features without explicit T-junction detection and interpretation. Int. J. Comput. Vis. **104**(1), 38–68 (2013)
3. Ng, R., Levoy, M., Brédif, M., Duval, G., Horowitz, M., Hanrahan, P.: Light field photography with a hand-held plenoptic camera. Computer Science Technical report (2005)
4. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. In: NIPS (2014)
5. Eigen, D., Fergus, R.: Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: ICCV (2015)
6. 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 (2015)
7. Wang, P., Shen, X., Lin, Z., Cohen, S., Price, B., Yuille, A.L.: Towards unified depth and semantic prediction from a single image. In: CVPR (2015)
8. Chakrabarti, A., Shao, J., Shakhnarovich, G.: Depth from a single image by harmonizing overcomplete local network predictions. In: NIPS (2016)
9. Ummenhofer, B., et al.: DeMoN: depth and motion network for learning monocular stereo. arXiv preprint arXiv:1612.02401 (2016)
10. Pentland, A.P.: A new sense for depth of field. IEEE Trans. PAMI **9**(4), 523–531 (1987)
11. Levin, A., Fergus, R., Durand, F., Freeman, W.T.: Image and depth from a conventional camera with a coded aperture. ACM Trans. Graph. **26**, 70 (2007)
12. Trouvé, P., Champagnat, F., Le Besnerais, G., Sabater, J., Avignon, T., Idier, J.: Passive depth estimation using chromatic aberration and a depth from defocus approach. Appl. Opt. **52**(29), 7152–7164 (2013)
13. Martinello, M., Favaro, P.: Single image blind deconvolution with higher-order texture statistics. In: Cremers, D., Magnor, M., Oswald, M.R., Zelnik-Manor, L. (eds.) Video Processing and Computational Video. LNCS, vol. 7082, pp. 124–151. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-24870-2_6
14. Sellent, A., Favaro, P.: Which side of the focal plane are you on? In: ICCP (2014)

15. Zhuo, S., Sim, T.: Defocus map estimation from a single image. Pattern Recognit. **44**, 1852–1858 (2011)
16. Carvalho, M., Saux, B.L., Trouvé-Peloux, P., Almansa, A., Champagnat, F.: On regression losses for deep depth estimation. In: ICIP (2018, to appear)
17. Chen, W., Fu, Z., Yang, D., Deng, J.: Single-image depth perception in the wild. In: NIPS, pp. 730–738 (2016)
18. Saxena, A., Chung, S.H., Ng, A.Y.: Learning depth from single monocular images. In: NIPS (2006)
19. Cao, Y., Wu, Z., Shen, C.: Estimating depth from monocular images as classification using deep fully convolutional residual networks. IEEE Trans. Circuits Syst. Video Technol. **28**(11), 3174–3182 (2017)
20. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR, pp. 770–778 (2016)
21. Xu, D., Ricci, E., Ouyang, W., Wang, X., Sebe, N.: Multi-scale continuous CRFs as sequential deep networks for monocular depth estimation. arXiv preprint arXiv:1704.02157 (2017)
22. Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., Navab, N.: Deeper depth prediction with fully convolutional residual networks. In: 2016 Fourth International Conference on 3D Vision (3DV), pp. 239–248. IEEE (2016)
23. Jung, H., Kim, Y., Min, D., Oh, C., Sohn, K.: Depth prediction from a single image with conditional adversarial networks. In: ICIP (2017)
24. Goodfellow, I., et al.: Generative adversarial nets. In: NIPS (2014)
25. Godard, C., Mac Aodha, O., Brostow, G.J.: Unsupervised monocular depth estimation with left-right consistency. arXiv preprint arXiv:1609.03677 (2016)
26. Garg, R., B.G., V.K., Carneiro, G., Reid, I.: Unsupervised CNN for single view depth estimation: geometry to the rescue. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9912, pp. 740–756. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46484-8_45
27. Kendall, A., Gal, Y.: What uncertainties do we need in Bayesian deep learning for computer vision? arXiv preprint arXiv:1703.04977 (2017)
28. Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: CVPR (2017)
29. Levin, A., Weiss, Y., Durand, F., Freeman, W.: Understanding and evaluating blind deconvolution algorithms. In: CVPR, pp. 1–8 (2009)
30. Veeraraghavan, A., Raskar, R., Agrawal, A., Mohan, A., Tumblin, J.: Dappled photography: mask enhanced cameras for heterodyned light fields and coded aperture refocusing. ACM Trans. Graph. **26**, 69 (2007)
31. Chakrabarti, A., Zickler, T.: Depth and deblurring from a spectrally-varying depth-of-field. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7576, pp. 648–661. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33715-4_47
32. Hazirbas, C., Leal-Taixé, L., Cremers, D.: Deep depth from focus. arxiv preprint arXiv:1704.01085, April 2017
33. Guichard, F., Nguyen, H.P., Tessières, R., Pyanet, M., Tarchouna, I., Cao, F.: Extended depth-of-field using sharpness transport across color channels. In: Rodricks, B.G., Süsstrunk, S.E. (eds.) IS&T/SPIE Electronic Imaging, International Society for Optics and Photonics, pp. 72500N–72500N-12, January 2009
34. Delbracio, M., Musé, P., Almansa, A., Morel, J.: The non-parametric sub-pixel local point spread function estimation is a well posed problem. Int. J. Comput. Vis. **96**, 175–194 (2012)

35. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from RGBD images. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7576, pp. 746–760. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33715-4_54
36. Srinivasan, P.P., Garg, R., Wadhwa, N., Ng, R., Barron, J.T.: Aperture supervision for monocular depth estimation. arXiv preprint arXiv:1711.07933 (2016)
37. Anwar, S., Hayder, Z., Porikli, F.: Depth estimation and blur removal from a single out-of-focus image. In: BMVC (2017)
38. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
39. Srivastava, N., Hinton, G.E., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. **15**(1), 1929–1958 (2014)
40. Hasinoff, S.W., Kutulakos, K.N.: A layer-based restoration framework for variable-aperture photography. In: 2007 IEEE 11th International Conference on Computer Vision, pp. 1–8, October 2007
41. Trouvé, P., Champagnat, F., Le Besnerais, G., Idier, J.: Single image local blur identification. In: IEEE ICIP (2011)
42. Kendall, A., Badrinarayanan, V., Cipolla, R.: Bayesian SegNet: model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. arXiv preprint arXiv:1511.02680 (2015)
43. Gal, Y., Ghahramani, Z.: Dropout as a Bayesian approximation: representing model uncertainty in deep learning. In: ICML, pp. 1050–1059 (2016)