NATIONAL UNIVERSITY FOR SCIENCE AND TECHNOLOGY POLITEHNICA BUCURESTI

Doctoral School of Electronics, Telecommunications and Information Technology

Decision No. 118 **from** 26/10/2023

Ph.D. THESIS SUMMARY

Saqib Nazir

OBŢINEREA HĂRŢILOR DE ADÂNCIME DIN IMAGINI DEFOCALIZATE FOLOSIND REŢELE NEURALE PROFUNDE

DEEP DEPTH FROM DEFOCUS FOR NEAR RANGE AND IN-SITU 3D EXPLORATION

THESIS COMMITTEE

Prof. Dr. Ing. Gheorghe BREZEANU UNSTPB	President
Prof. Dr. Ing. Daniela COLŢUC UNSTPB	PhD Supervisor
Prof. Dr. Ing. Víctor Manuel Brea SANCHEZ University of Santiago de Compostela (USC), Spain	Referee
Dr. Ing. Habil. Miguel HEREDIA CONDE University of Siegen, Germany	Referee
Prof. Dr. Ing. Mihai DATCU UNSTPB	Referee

BUCHAREST 2021

Abstract

An image taken with a conventional camera is a 2D projection of a 3D scene. During the imaging process, the information of the 3rd dimension i.e., the scene depth is lost. However, it can be recovered by computation, from the set of visual cues present in the images. This has created the premises to perceive the world in 3D by means of conventional cameras.

The thesis addresses the challenge of monocular depth estimation using a single defocused image, a pivotal task in computer vision given the wide area of applications ranging from robotics to self-driving cars. The recent advances in deep learning has revolutionized the field of computer vision and in particular, depth estimation. However, prior deep learning-based techniques often neglect the potential of defocus blur, which is an important cue for depth estimation. The existing solutions employ multiple images or focal stacks of the same scene and rarely single images. To fill in this gap, we propose a novel architecture called 2HDED:NET, that addresses both depth estimation and image deblurring from a single defocused image.

Due to the absence of datasets containing naturally defocused images and depth ground truth, networks like 2HDED:NET are typically trained on synthetic data, a fact that reduces their performances on real data. In order to train 2HDED:NET to its full potential, we proposed a new dataset called iDFD containing naturally defocused images, double annotated with the all-in-focus image, and the Time of Flight depth map.

Finally, the thesis explores the promising field of self-supervised learning, by converting 2HDED:NET to defocus map estimation in the absence of ground truth depth for training. To accomplish this, 2HDED:NET is enhanced with a defocus simulation module that reconstructs the defocused image from the all-in-focus one and the estimated defocus map. The model's proficiency in defocus map estimation is on par with that of state-of-the-art supervised models that use multiple images. Comprehensive experiments conducted on various real or synthetic datasets validated the efficacy of the proposed approaches for depth or defocus map estimation in various settings, encompassing indoor and outdoor environments.

Table of contents

Chapter 1

Introduction

1.1 Presentation of the field of the doctoral thesis

Computer vision (CV) is a field of Artificial Intelligence (AI), that enables computers to extract meaningful information from digital images. This thesis falls into the category of early CV, addressing the challenging and fundamental problem, called Monocular Depth Estimation (MDE). With the Deep Learning (DL) models, it is possible to extract deep features from a single-shot RGB image, which can provide better results than traditional analytical methods.

1.2 Scope of the doctoral thesis

MDE has gained significant improvements with the development of Deep Neural Networks (DNNs) in recent years, but the importance of defocus blur is yet to be exploited. Hence, the scope of this thesis encompasses the study and development of a novel approach for depth estimation using defocus blur as a cue. The primary focus is on investigating the relationship between defocus blur and depth information in monocular images. A DNN called 2HDED:NET is designed to perform depth estimation from defocus blur while simultaneously addressing image deblurring. The thesis involves the development of a dataset called iDFD, consisting of real-world defocus images, their All-in-Focus (AiF) counterparts, and corresponding depth maps. This dataset serves as a valuable resource for training and evaluating the proposed DNN model. Furthermore, we enhanced 2HDED:NET, to enable self-supervised learning for defocus map estimation. The network is designed to learn and estimate defocus maps in a completely self-supervised manner, removing the need for labeled annotation or explicit supervision.

1.3 Content of the doctoral thesis

The second chapter presents the recent developments in MDE using DL. There are various choices for the architectures, loss functions, smoothing regularization, and experimental setups proposed in the literature. Hence it is difficult to establish their respective influence on the performances. We made a comparison of different smoothing regularizations proposed in the literature in both supervised and self-supervised methods. In addition to this, we used a smoothing regularization term used in a self-supervised way to work in a supervised manner and show that the modified technique shows considerable performance by accurately producing edges of the objects. Parts of this chapter were published in [12].

The third chapter focuses on exploring the defocus blur as a cue for depth estimation. Depth estimation and image deblurring are two fundamental and closely related tasks. Performing any of them by relying on a single image is an ill-posed problem. Despite this, most of the existing models treat them separately. In this chapter, we proposed the Two-headed Depth Estimation and Deblurring Network (2HDED:NET), which extends a conventional Depth from Defocus (DFD) network with a deblurring branch that shares the same encoder as the depth branch. Parts of this chapter were published in [14, 15].

Many datasets have been proposed in the literature with the aim of depth estimation, with NYU-Depth V2 and KiTTi being the most famous ones. But when it comes to DFD there is no such dataset with the natural defocus images. Hence most of the previous methods train their DNN in synthetic datasets. To overcome the limitation of a dataset with the natural defocus blur, we proposed Indoor Depth from Defocus (iDFD), a Depth And Defocus Annotated dataset, which contains naturally defocused, AiF images and dense depth maps of indoor environments. Parts of this chapter were published in [13].

In this chapter, we propose a self-supervised learning model for Defocus Map Estimation (DME) from a single defocused image. Our method is based on a recently proposed DNN called 2HDED:NET that we complete with a defocus simulation module. We show that our self-supervised network circumvents the need for Ground Truth (GT) defocus or depth maps. In addition to the DME, our network reconstructs the AiF image through supervised learning. We test the network on synthetic and realistic benchmarks and demonstrate that it is an effective solution for DME and image deblurring when a single defocused image is available.

Chapter 2

Monocular Depth Estimation using Deep Learning

A number of different deep architectures have been proposed in recent years, the choice of appropriate architecture, loss function, hyper-parameters, and pre-trained models is important. In this chapter, we discussed the best-performing models along with their choices of networks, etc. Moreover, we analyzed different smoothing regularization terms and selected the most efficient ones used in the previous studies.

2.1 Smoothing Regularization

The smoothing regularization is employed for smoothing the homogeneous areas of the objects in a scene without degrading the edges of a predicted depth. Considering the location of the abrupt edges in the natural images is unknown, they should be identified at the same time as the object is being reconstructed. In order to encourage smoothing within a flat region and discourage smoothing across edges, different solutions have been proposed in the previous studies [5, 18, 9, 20, 6, 7]. This chapter makes a comparison between various regularization terms used either in supervised or self-supervised learning methods. In addition to this, the regularization term currently used in self-supervised methods has been modified, such as working in a supervised manner. The experimental results on NYU-Depth v2 have shown that the regularization based on L1 norm of the gradient is the best and the self-supervised modified one outperforms the rest. Finally, rather than relying on common evaluation metrics, an additional accuracy measure based on the Steerable Pyramid and Kullback-Leibler divergence (KLD) is used for edge accuracy of estimated depths that are more sensitive to positional errors of the edges.

2.1.1 Metrics for Edge Sharpness

The basic functions of the steerable wavelet transform are the K_{th} order directional derivatives, which come in multiple scales and 4 orientations. An example decomposition

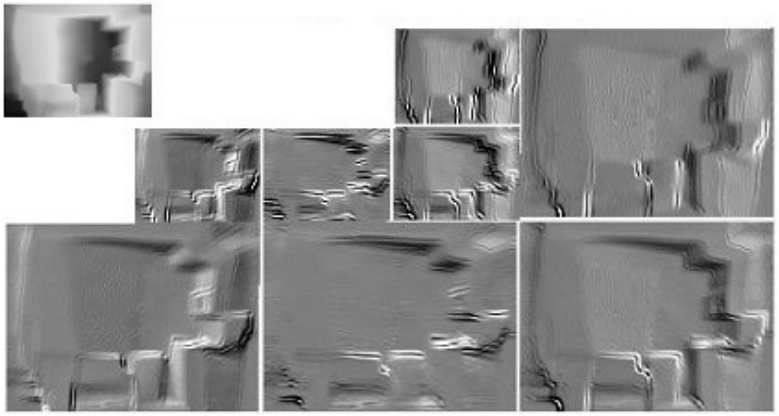


Figure 2.1 Steerable pyramid decomposition.

of the estimated depth map is shown in Figure 2.1. This particular steerable pyramid contains 4 orientation subbands, at 2 scales. The reason for selecting the first two scales is that they contain the highest spatial frequency of the image. Given that the edge degradation caused by smoothing is primarily noticeable in the high-frequency domain and taking into account the dimensions of the test images, the generation of lower scales was not pursued due to the insignificant impact they would have had on the evaluation of the KLD.

To evaluate the edges, both the GT and the estimated depth are decomposed on 2 scales and 4 orientations, resulting in 8 sub-bands, then the KLD is calculated on sub-bands. To estimate the KLD, histograms of each sub-band are created, and the divergence of corresponding sub-bands is calculated as described in [19]:

$$d(h_m||h) = \sum_{i=1}^{L} h_m(i) \log \frac{h_m(i)}{h(i)} \tag{2.1}$$

where $h_m(i)$ and $h(i)$ are normalized heights of i_{th} histograms of GT depths and estimated depths, respectively, and L is the number of the bins in the histograms.

Finally, the global KLD between the estimated and GT depths is obtained as:

$$D = \log_2(1 + \frac{1}{D_o} \sum_{K=1}^{K} |d^k(h_m^k||h^k) \tag{2.2}$$

where K is the number of subbands, and D_o is the constant used to control the scale of the distortion measure, in our case $k = 8$ and $D_o = 10$.

2.2 Experimental Results

Table 2.1 shows the results obtained for 5 different regularizations and for the case without any regularization. It can be seen that the worst results are obtained without the smoothing regularization, which demonstrates once a more the usefulness of introducing

Table 2.1 Quantitavie comparison with various smoothing regularisations. For RMSE, REL, Log 10, and KLD, lower is better. For δ_1, δ_2, and δ_3 higher is better.

Method	***RMSE***	***REL***	***Log 10***	δ_1	δ_2	δ_3	***KLD***
Non-Regularization	0.619	0.159	0.064	0.786	0.950	0.986	9.0091e+03
Hu[9]	**0.555**	0.126	0.054	0.841	0.967	0.991	7.2385e+03
Eigen[5]	0.598	0.143	0.060	0.812	0.956	0.988	7.3216e+03
Ummenhofer[18]	0.582	0.134	0.058	0.822	0.963	0.991	7.2817e+03
Xian[20]	0.559	**0.125**	**0.054**	**0.844**	**0.968**	**0.991**	**7.1009e+03**
Godard[7] (Modified)	0.575	0.142	0.057	0.825	0.961	0.988	7.1929e+03

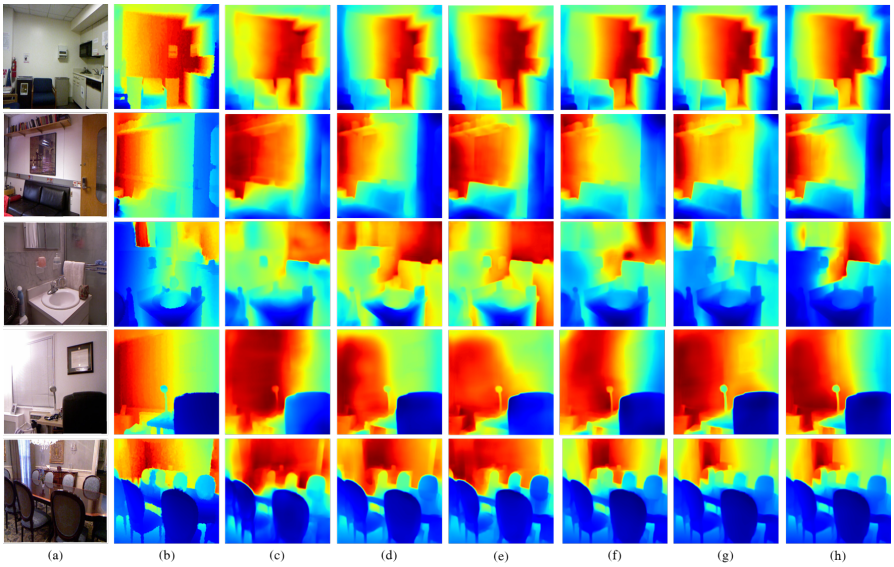


Figure 2.2 Estimated depth maps for 6 scenes and 5 different smoothing regularizations. From the first to the last row; (a) input RGB images, (b) GT depth map, (c) results without regularization term, (d) Hu [9], (e) Eigen [5], (f) Ummenhofer [18], (g) Xian [20], and (h) Godard [7] (modified) loss function.

such regularization. The best results are obtained for Hu [9] and Xian [20], the later one slightly in advantage except for RMSE. Godard [7] overcomes Eigen [5], although it comes from the self-supervised methods.

The results of *KLD* for the edge accuracy are shown on the last column of Table2.1. Although supervised methods produced good performance on common evaluation metrics, their accuracy is not appreciable on edge accuracy except Xian [20], which is by far the best one. Surprisingly, the second best results are provided by the modified solution coming from the self-supervised methods. The explanation is in using the same L1 norm. It has been known that the L1 norm favors sparser errors. In smoothing regularization, L1 norm is applied to the gradient of images, which reveals the sparsity.

Figure 2.2 shows the depth maps estimated by using various smoothing regularizations, and their corresponding RGB and GT images. It can be seen that the results without the regularization term suffer from heavy distortion of shapes. Although Hu and Eigen [9, 5] were able to produce images with clear boundaries, they present however several incorrect soft edges, for example, the boundaries of the light lamp on the desk in the case of Hu [9] and objects on the sink in the case of Eigen [5]. The modified regularization term in Godard [7] shows considerable performance by accurately producing edges of the objects and minute structures, such as objects around the sink and light lamp on the desk. Also in the image with the sofa, clear boundaries can be seen. Here, Xian [20] was unable to detect the accurate boundaries of this image.

Chapter 3

Depth Estimation from Defocused Images

This chapter presents our proposed DNN framework for jointly addressing DFD and image deblurring tasks using a single defocused image. While these tasks are related, existing models typically treat them as separate problems. Recent DNN-based methods that aim to solve these tasks simultaneously first estimate the depth or defocus map and then reconstruct the focused image based on this estimation [1]. In our approach, a Two-headed Depth Estimation and Deblurring Network (2HDED:NET) is introduced, which extends a conventional DFD network by incorporating a deblurring branch that shares the same encoder as the depth branch.

3.1 2HDED:NET Architecture

Figure 3.1 depicts the architecture of 2HDED:NET. Given a single defocused image I, the goal of our network is to estimate the depth map $\widehat{I}^{depth}$ and to restore the AiF image $\widehat{I}^{aif}$. As shown in Figure 3.1, 2HDED:NET consists of one encoder and two decoders that output the depth map and AiF image in parallel. By utilizing the features learned by the same encoder, both heads can mutually benefit from each other. 2HDED:NET is a supervised method that requires the GT depth as well as the AiF images for training.

For the encoder network, the DenseNet-121 [2] is used. As its name suggests, DenseNet consists of densely connected layers. Similar to [2], the max-pooling layer is replaced with a 4×4 convolutional layer to reduce resolution while increasing the number of the feature channel maps. The skip connections are used between the encoder and decoder parts to simplify learning. The encoder helps to obtain multi-resolution features from the input image, which are useful for the two tasks that 2HDED:NET performs.

The Depth Estimation Decoder (DED) is inspired by [2]. It consists of 5 decoding layers, each with 4×4 convolution that increases the resolution of the feature map,

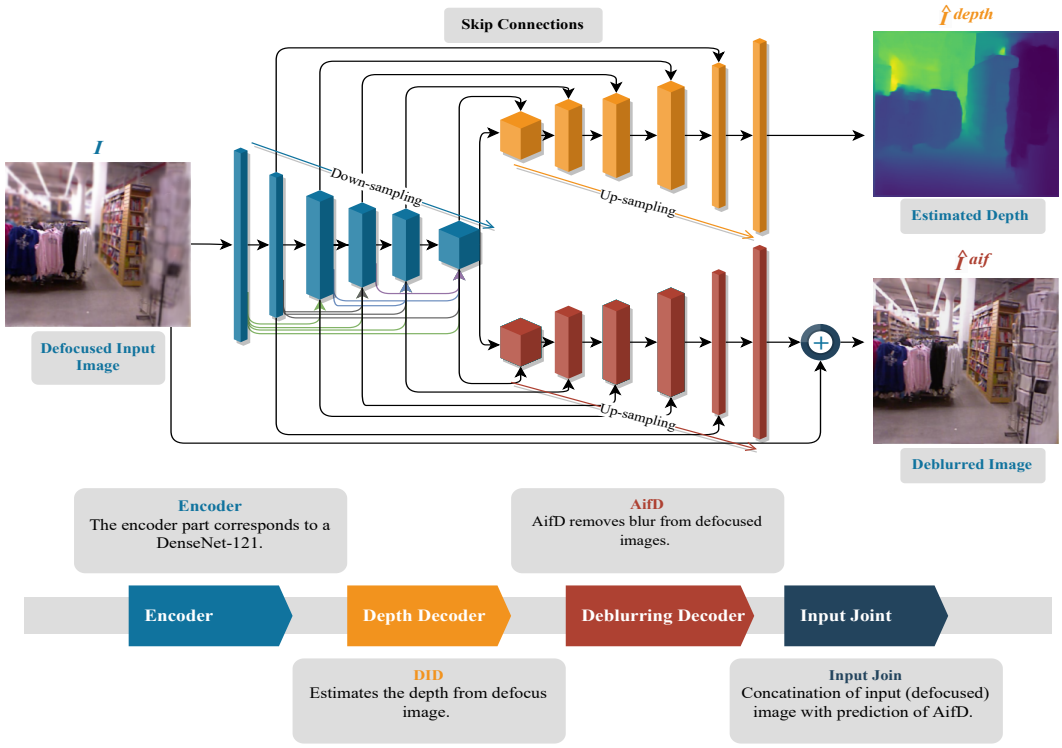


Figure 3.1 2HDED:NET architecture consists of one encoder and two decoders that work in parallel. The upper Head estimates the depth map and the lower one the AiF image. The network is fed in with defocused RGB images.

followed by a 3×3 convolution that reduces the aliasing effect of upsampling. Batch normalization and ReLU functions are included after each convolutional layer to make learning more stable and to allow the representation of nonlinearities.

The deblurring decoder is referred to as the AiF decoder (AifD). Unlike DED, the output of AifD is a three-channel RGB image. An input joint layer is used to aggregate the defocused input image with the output of AifD for the final prediction. The content of the defocused image and the corresponding prediction from AifD are embedded in the input joint layer, giving this head more detailed guidance for learning deblurring. Unlike methods that use pipeline processing, where the depth or defocus map is first predicted and then the AiF image is recovered, our deblurring head is not based on such estimates, avoiding reliance on insufficient depth maps in some cases.

An important feature of the proposed solution is that once 2HDED:NET is fully trained, a task can still be performed when the other head is removed, e.g. DFD can be performed without the AifD head and vice versa.

The training of the 2HDED:NET is supervised simultaneously by GT depth maps and AiF images. To consider this dual information, a loss function is proposed with two terms, one that accounts for the depth loss and another for the deblurred image. These two components are balanced to have approximately equal contributions.

Most of the DL methods proposed for depth estimation have been trained with pixel-wise regression-based loss functions calculated as the mean of absolute differences ($L1$ norm), squared differences ($L2$ norm), or combinations of them [2].

The $L1$ norm is resorted to as the loss function for depth estimation, recognized for its capability to estimate sparse solutions as observed in the case of depth maps [2, 14]:

$$L_1^{Depth} = \frac{1}{n}\sum_{i=1}^{n} |\widehat{I}_i^{depth} - I_i^{depth}| \tag{3.1}$$

where $\widehat{I}^{depth}$ is the estimated depth, I^{depth} the GT, i is the current pixel and n is the number of pixels.

Often, this loss is complemented by a smoothing regularization term that has the role of removing the low amplitude structures in the depth map while sharpening the main edges [7, 8, 20]. In the case of the proposed network, the depth accuracy is improved by combining $L1$ norm with the smoothing term commonly used in supervised learning and defined as [20]:

$$L_{grad} = \frac{1}{n}\sum_{i} |\Delta_x R_i| + |\Delta_y R_i| \tag{3.2}$$

where $R_i = \widehat{I}_i^{depth} - I_i^{depth}$ and Δ_x and Δ_y are the spatial derivatives with respect to the x-axis and y-axis. As a result, the overall depth loss function is defined as (3):

$$L_{depth} = L_1^{Depth} + \mu L_{grad} \tag{3.3}$$

where μ is a weighting coefficient set to 0.001.

Various loss functions have been proposed to train the DNNs for image deblurring. Pixel-wise content loss functions like $L1$ and $L2$ norm are the most common. For the training of 2HDED:NET, the $L1$ norm and Charbonnier loss function [3], which is the smoothed version of $L1$, were tested. Charbonnier loss is calculated as a squared error between the estimated deblurred image $\widehat{I}^{aif}$ and the GT AiF image I^{aif}:

$$L_{charb} = \frac{1}{n}\sum_{i=1}^{W}\sum_{j=1}^{H} \sqrt{(\widehat{I}_{i,j}^{aif} - I_{i,j}^{aif})^2 + \varepsilon^2} \tag{3.4}$$

where ε is a hyper-parameter set to $1e-3$. This hyper-parameter acts as a pseudo-Huber loss and smooths the errors smaller than ε.

The loss function defined either as Charbonnier or $L1$ norm, is improved by requiring a high SSIM. This results in adding the regularization term:

$$L_{SSIM} = 1 - SSIM(\widehat{I}_{i,j}^{aif}, I_{i,j}^{aif}) \tag{3.5}$$

which makes the complete deblurring loss function to be:

$$L_{deblur} = L_{charb} + \Psi L_{SSIM} \tag{3.6}$$

where Ψ is a weight set to 4.

With the depth and deblurring losses defined as in Equation3.3 and 3.6, total loss for 2HDED:NET training is the followng:

$$L_{2HDED} = L_{depth} + \lambda L_{deblur} \tag{3.7}$$

3.2 Experimental Results

The synthetically defocused images have been used by many recent works [1, 2] dedicated either to depth inference or image restoration. For the experiments, we use NYU-Depth-V2, Make3D, DFD, and DIOD benchmarks. Table 3.1 presents in the left half, results for depth estimation obtained with networks trained on NYU-Depth-V2 dataset. For the NYU-Depth-V2 dataset, the best accuracy in terms of RMSE is obtained by Carvalho et al. [2]. From the same category of networks using defocused images, there is [1]. On average, the depth maps accuracy of [1] is worse by 0.2 in RMSE compared with the best result in [2]. 2HDED:NET is at half way between [2] and [1] with a RMSE of 0.244. In the category of networks handling both depth and deblurred images, our 2HDED:NET is the best in all metrics.

As already stated by [1, 2], employing out-of-focus images, as opposed to AiF images, leads to significant improvement in depth estimation accuracy. Qualitative results on NYU-Depth-V2 dataset for both AiF and out-of-focus cases are depicted in Figure 3.2. In all examples, the improvement is evident when out-of-focus images are used for training. In the case of AiF images, although the 2HDED:NET is able to estimate the closer and distant regions, the scene's content remains indiscernible. The content becomes evident only when the depth maps are estimated from out-of-focus images.

In the evaluation of 2HDED:NET, particular emphasis is placed on comparing it with the network proposed by Anwar et al. [1]. Although both networks provide depth maps

Table 3.1 Comparison of 2HDED:Net with SoA methods for depth estimation and image deblurring on NYU-Depth-V2 datasets.

			Depth Estimation		Deblurring	
Method	Depth	Deblur	*RMSE* ↓	*Abs. rel* ↓	*PSNR* ↑	*SSIM* ↑
Carvalho et al. [2]	✓	×	**0.144**	0.036	–	–
Anwar et al. [1]	✓	✓	0.347	0.094	34.21	–
2HDED:Net	✓	✓	0.244	**0.029**	**34.85**	0.99

and deblurred images, Anwar's network utilizes a pipeline processing approach, making it a suitable point of comparison with our network.

In Figure 3.3, image deblurring results on NYU-Depth-V2 dataset are given, restored by both [1] and 2HDED:NET. Our method achieves a PSNR of 34.85 dB, which is almost 0.7 dB higher than that of [1] for the same image. The blur removal can be well observed in the areas delimited by the red rectangles: the light on the ceiling and the edges of the furniture.

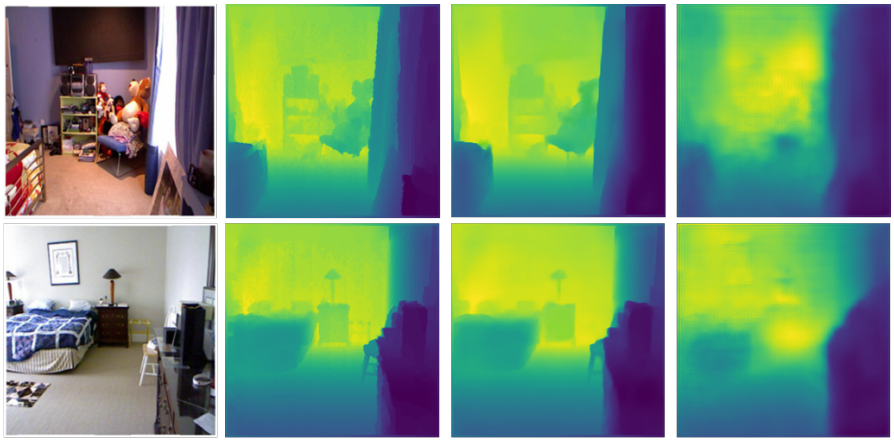


Figure 3.2 Examples of depth estimation with 2HDED:NET (deblurring head ablated) from NYU-Depth-V2 benchmarks after training with defocused or AiF images. (a) AiF RGB images, (b) GT depth, (c) Depth estimated with defocused images, (d) Depth estimated with AiF images.

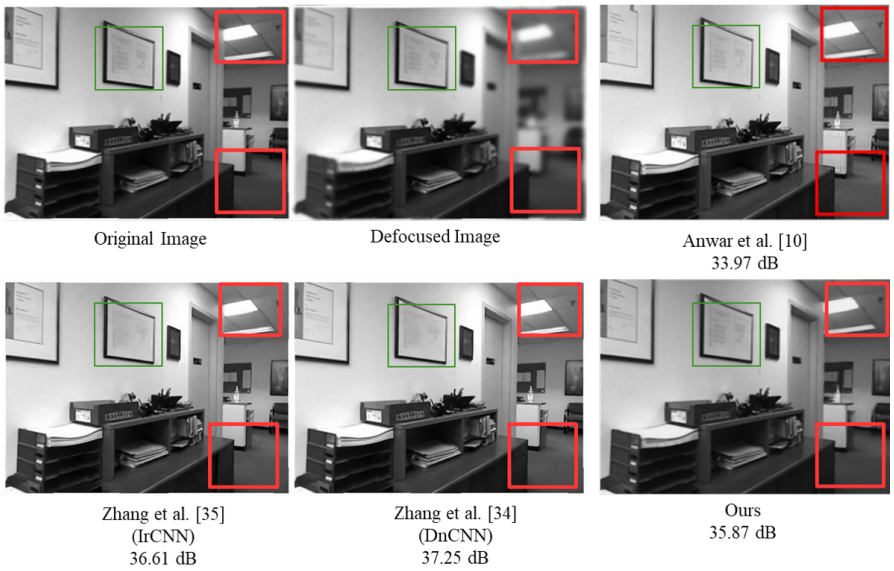


Figure 3.3 Comparison of 2HDED:NET with the pipeline solution of Anwar et al. [1], and two general methods for image restoration [21, 22]: an example from NYU-Depth-V2 dataset.

Chapter 4

iDFD: A dataset for Depth Estimation and Defocus Deblurring

In this chapter, a dataset is proposed for depth estimation and defocus deblurring based on naturally defocused images. The iDFD dataset proposed in this chapter is a large-scale dataset with 764 image triples consisting of AiF images, defocused counterparts, and corresponding depth maps.

In recent years, numerous DL-based techniques for MDE have emerged. Among these, supervised methods have consistently demonstrated the most impressive results. The success of supervised DL methods however depends on the existence of large diverse training benchmarks such as KITTI, Cityscapes, or Make3D that contain RGB images and their corresponding GT consisting in either dense depth maps or cloud of points.

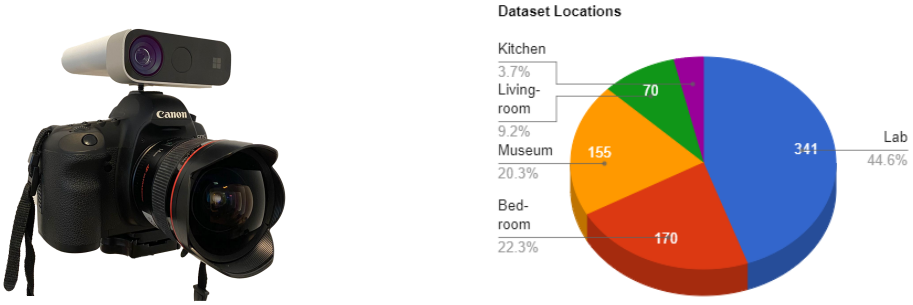
4.1 iDFD Dataset for DFD and Image Deblurring

In the previous chapter, a MTL network is proposed called 2HDED:NET [14], where DFD and deblurring support each other to better accomplish these tasks. 2HDED:NET is trained on a synthetically defocused dataset without confirming its effectiveness on naturally defocused images because of the lack of such dataset [14].

A consistent dataset with defocused images annotated for depth and defocus would boost the supervised DL methods for MDE and image deblurring and improve their accuracy. To fill in this gap, iDFD, a Depth, and Defocus Annotated dataset is introduced, with naturally defocused images and corresponding depth and AiF GT. iDFD is a large-scale dataset with various indoor scenes. The dataset is acquired with a MS-Kinect camera and a Digital Single-Lens Reflex (DSLR) camera tightly coupled to each other. The DSLR camera is adopted to capture the naturally defocused images since the RGB camera of MS-Kinect does not have the facility to work with different apertures. Table 4.1 shows an overview of iDFD dataset. The dataset contains a variety of indoor scenes captured in 5 different environments: bedroom, living room, labs, kitchen, and museum.

Table 4.1 iDFD Overview.

Sensors	Scenes	Range	Data	Total no. Images	Depth Maps	Out-of-Focus DSLR Settings
DLSR-Nikon/ MS-KINECT	Indoor	$0-10m$	RGB/ Depth	764	Raw/ In-painted	Lens: 14mm $f/2.8$



(a) (b)

Figure 4.1 (a). Experimental setup: DSLR-Nikon camera is coupled with MS-Kinect camera using a Camera Shoe Mount Adapter. Two cameras are fixed on a tripod to avoid motion. The DSLR is used to capture AiF and Defocus (RGB) images and Kinect is used to capture the GT depth for RGB images. (b). Data acquisition locations and the total number of images per location.

In total, 764 scenes are captured, most of them coming from the labs. Figure 4.1b shows the distribution of scenes over the 5 environments. Some scenes, such as those from bedrooms and living rooms, are much simpler as content than the scenes coming from labs and museums. The goal has been to acquire three images for each scene: defocused, AiF, and depth.

A DSLR-Nikon camera with a $14mm$ lens at two different aperture sizes is used to capture AiF and defocused images. For AiF images a narrow aperture with $f/10$ is used, and for the out-of-focus ones a wider aperture with $f/2.8$. The wide aperture creates a shallow depth of field resulting in a blurred image.

To acquire the GT depth, the MS-Kienct camera in the Narrow Field-of-View (NFOV) depth mode is used with an operating range of $0.5-3.8m$, and the maximum Field of Interest (FoI) of $75^\circ \times 65^\circ$. The depth can also be provided outside the operating range, but it depends entirely on the reflectivity of the objects. In the selection of scenes, areas where depth information might not be accurately captured, such as shadows, were avoided. The experimental setup is shown in Figure 4.1a. The Kinect camera is mounted on a DSLR-Nikon using a shoe mount adapter, and our experimental setup is similar to that of Carvalho et al. [2]. For pre-processing the data, the procedure of Qiu et al. [16] is followed, consisting of denoising, inpainting (optional), image registration, normalization, and cropping.

4.2 Experimental Results

In this section, the iDFD dataset is evaluated on 2HDED:NET proposed in the previous chapter. To emphasize the importance of using real data, the network is retrained under the same conditions on the NYU-Depth-V2 dataset, supplemented with synthetically defocused images. The results obtained on the two benchmarks are compared in terms of depth accuracy and deblurring.

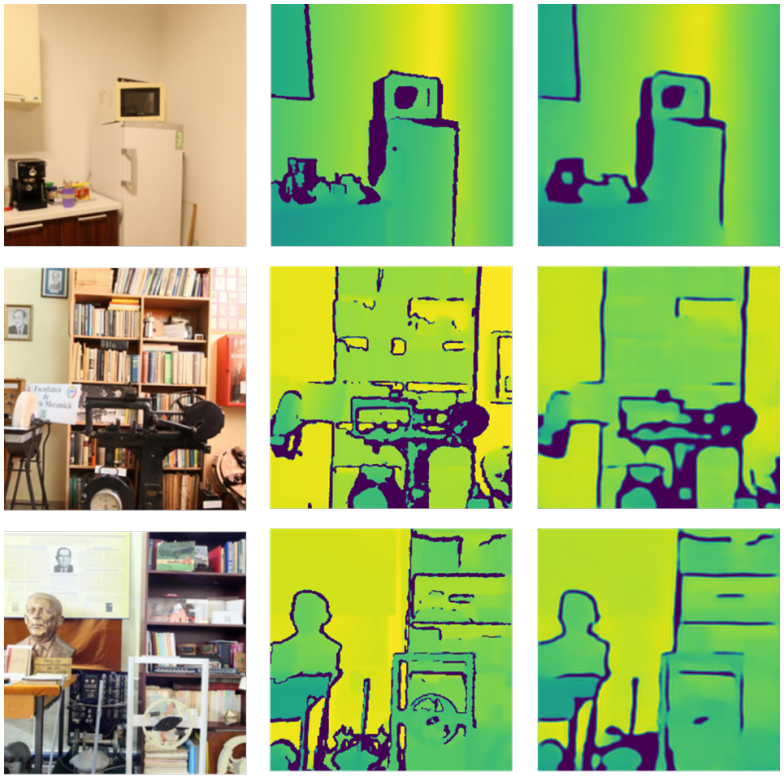
Table 4.2 presents the errors in estimating the depth for both NYU-Depth-V2 and iDFD benchmarks. When training on NYU-Depth-V2 dataset, the RMSE is lower than that of the raw depths from the iDFD dataset. The figures show better accuracy when in-painted depth is used. These results must be taken under the reserve of a GT which is not 100% measured. The worst results are obtained for the network trained on NYU-Depth-V2 and tested on iDFD dataset. Figures 4.2 and 4.3 depict three examples of depth maps obtained after training with raw depth and in-painted depth, respectively. The visual inspection shows that when raw images are used, the network tends to wipe small regions like the empty spaces on the shelves in the second example. These details are preserved by the depth maps obtained after training on in-painted depth. This suggests that in-painting should be performed rather before training then after training with raw depth.

Table 4.3 illustrates the deblurring results of 2HDED:NET on NYU-Depth-V2 and iDFD benchmarks. The 2HDED:NET has shown promising results for deblurring of synthetically defocused images from NYU-Depth-V2, where the PSNR is improved from 26.09 dB to 32.68 dB (Table 4.3). On iDFD naturally defocused images, 2HDED:NET is able to improve the PSNR from 25.83 dB to 36.25 dB, which means a gain of 10.43 dB, significantly higher than the 6.59 dB obtained in the case of NYU-Depth-V2. The initial PSNR of 25.83 dB has been calculated by taking as reference the AiF image obtained with aperture $f/10$.

Finally, the network trained on NYU-Depth-V2 is tested on iDFD. The deblurring has been more modest with a final PSNR of only 30.09 dB. This result shows that a network trained on a dataset with natural defocus and high quality RGB images can be more effective for image deblurring than the same network trained on a large synthetically defocused dataset, such as that obtained from NYU-Depth-V2.

Table 4.2 Depth Estimation by 2HDED:NET on NYU-Depth-V2 and iDFD benchmarks.

Training/Testing dataset	*RMSE* $\downarrow$	*Abs. rel* $\downarrow$	$\delta 1 \uparrow$	$\delta 2 \uparrow$	$\delta 3 \uparrow$
NYU-Depth-V2/NYU-Depth-V2	0.281	0.266	0.877	0.942	0.958
NYU-Depth-V2/iDFD	0.401	0.284	0.759	0.813	0.839
iDFD/iDFD (raw depth)	0.312	0.194	0.727	0.793	0.807
iDFD/iDFD (in-painted depth)	0.248	0.145	0.799	0.854	0.959



(a) (b) (c)

Figure 4.2 Results with 2HDED:NET trained on raw depth: a) RGB image, b) Raw GT depth, c) Estimated depth.

Figure 4.4 depicts two images with a complex content, restored by 2HDED:NET trained on iDFD dataset. Crops representing the microscope in the first image or the pillow and the picture in the second image, are magnified in order to show how details sinked into blur, and emerge after restoration with 2HDED:NET.

Table 4.3 Deblurring results obtained with 2HDED:NET trained or iDFD and NYU-Depth-V2 benchmarks.

Dataset	*PSNR* ↑	*SSIM* ↑
NYU-Depth-V2 after deblurring (training on NYU-Depth-V2)	32.68	0.91
NYU-Depth-V2 defocused	26.09	0.49
Gain	6.59	0.42
iDFD after deblurring (training on iDFD)	36.25	0.99
iDFD defocused	25.83	0.51
Gain	10.43	0.43
iDFD after deblurring (training on NYU-Depth-V2)	30.09	0.86

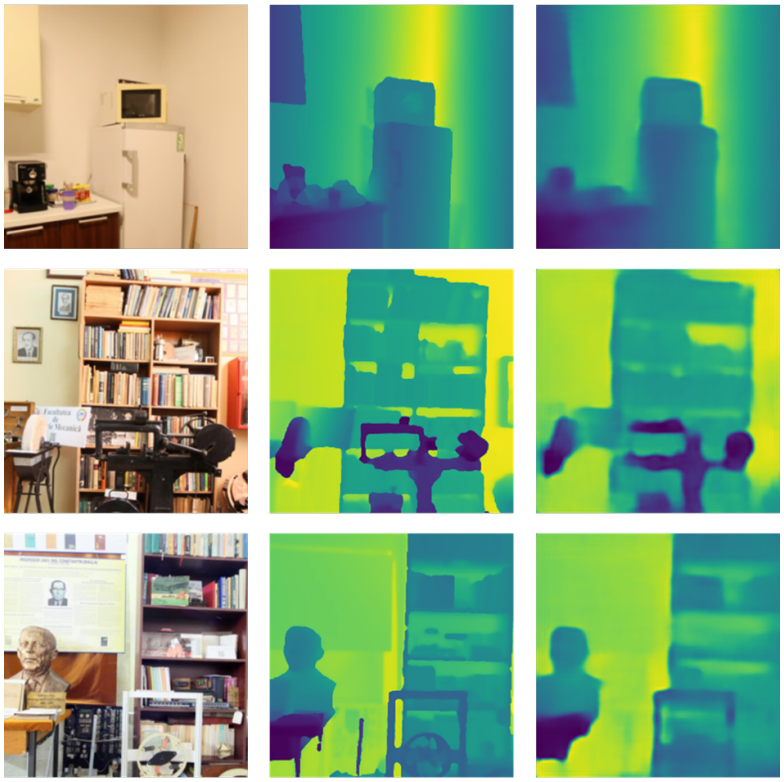


Figure 4.3 Results with 2HDED:NET trained on in-painted depth: a) RGB image, b) Depth GT, c) Estimated Depth.

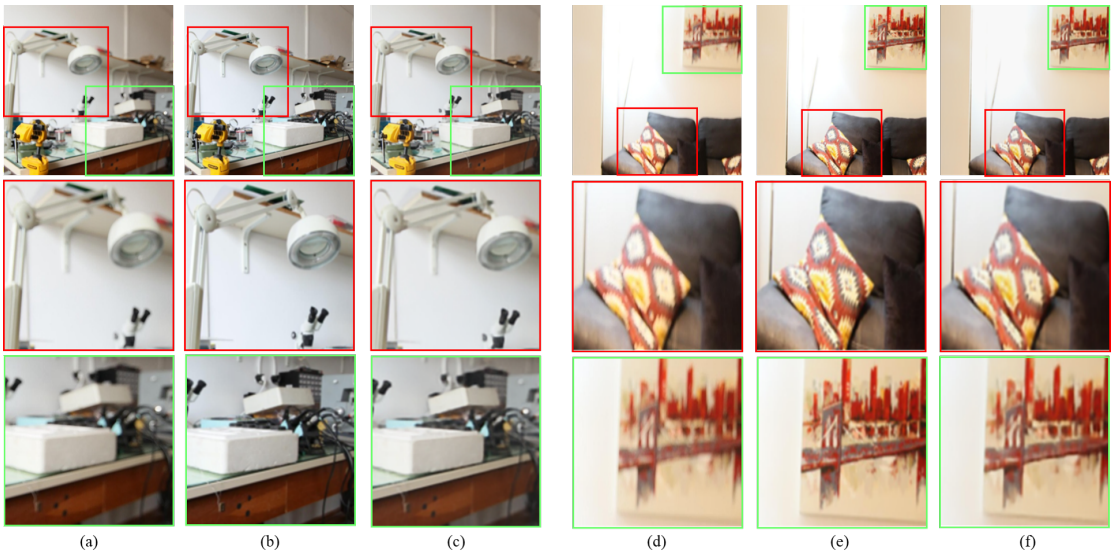


Figure 4.4 2HDED:NET results for image deblurring: from left to right in the first-row a) Defocus (input), b) AiF (GT), and c) Deblurred. Similarly, d) Defocus (input), e) AiF (GT), and f) Deblurred. The second row shows the zoomed-in regions to illustrate the deblurring results. Two examples are taken from the electronics lab and living room.

Chapter 5

Self-supervised Learning for Defocus Map Estimation

In this chapter, an end-to-end self-supervised DNN designed for Defocus Map Estimation (DME) from a single defocused RGB image is proposed. The proposed method is built upon the foundations of the 2HDED:NET introduced in Chapter 3, but the network is enhanced with a defocus simulation module that enables self-supervised training for DME. In addition to the defocus map, our network also reconstructs the AiF image using supervised learning. The network is tested on synthetic and real benchmarks, demonstrating its effectiveness for DME and image deblurring when a single defocused image is available.

5.1 Self-supervised 2HDED:NET

The entire training pipeline is summarized in Figure. 5.1. The 2HDED:NET is a MTL DNN proposed in Chapter 3, the network is designed to solve the DFD problem and reconstruct an AiF image from a single out-of-focus image in a supervised manner.

In this chapter, the utilization of defocus blur is capitalized upon. The approach involves training the 2HDED:NET as a self-supervised network. To accomplish this, the 2HDED:NET is augmented with a Defocus Simulation module, enabling the generation of a reblurred image during training. This specific reblurred image is derived from the estimated defocus map and the GT AiF image. To maintain consistency, the same parameters used in generating the training datasets are used when reblurring images during training.

To train the self-supervised 2HDED:NET, a loss function similar to that used in [7] is utilized, which incorporates the $L1$ norm to evaluate the absolute difference between the synthetically re-blurred RGB I^r and the input defocused image I^b. The loss function is:

$$L_{re} = \frac{1}{n}\sum_{t=1}^{n}\frac{\alpha}{2}(1 - SSIM(I^r - I^b)) + (1-\alpha)(|I^r - I^b|) \tag{5.1}$$

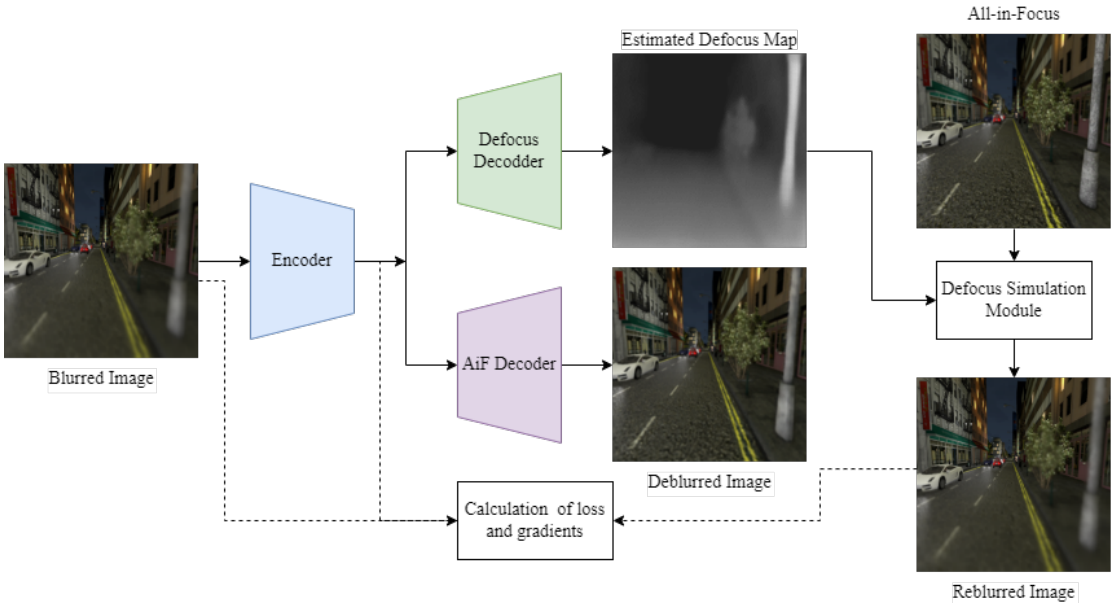


Figure 5.1 2HDED:NET architecture consists of one encoder and two decoders that work in parallel. The upper branch estimates the defocus map and the lower one the AiF image. The network is fed in with a defocused RGB image. The Defocus Simulation module reblur the images with the help of the Estimated Defocus map and AiF images during training.

where $|.|$ represents the $L1$ norm, and SSIM is the Structural Similarity Index. $\alpha = 0.85$ is a constant.

In order to prevent the predicted defocus map from drastic changes in homogeneous regions, the smoothing regularization constraint is introduced in the loss function. Similar smoothing regularization constrain have been used in almost all the supervised or self-supervised depth map estimation methods [17, 7, 12]. This constraint also increases the consistency between AiF and the estimated defocus map. The loss function is:

$$L_{smooth} = \frac{1}{n}\sum^{n} |\delta_x I^d| e^{-\beta \delta_x I^{AiF}} + |\delta_y I^d| e^{-\beta \delta_y I^{AiF}} \tag{5.2}$$

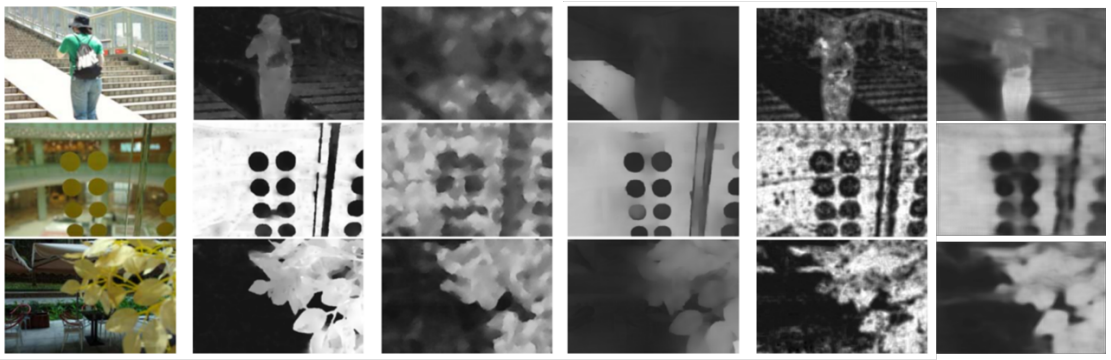
where I^d is the estimated defocus map, I^{AiF} is the GT AiF image, and δ is the gradient computed on the x and y axis respectively. β set to 2.5 is the scale factor for the sensitivity of the edges [17].

Various loss functions have been proposed to train the DNNs for image deblurring. To train 2HDED:NET, the Charbonnier loss function [15] is used in Chapter 3, which is the smoothed version of $L1$. Charbonnier loss is calculated as the squared error between the estimated deblurred image $\widehat{I}^{AiF}$ and the GT AiF image I^{AiF}:

$$L_{deblur} = \frac{1}{n}\sum_{i=1}^{W}\sum_{j=1}^{H} \sqrt{(\widehat{I}_{i,j}^{AiF} - I_{i,j}^{AiF})^2 + \varepsilon^2} \tag{5.3}$$

Table 5.1 Quantitative analysis for Defocus Maps Estimation and Image Deblurring on different datasets, best results are shown in bold letters.

	Trained on DFD & and tested on RTF dataset				Trained & tested on SYNDOF dataset	Trained & tested on iDFD dataset
	[4]	[10]	[11]	Ours	Ours	Ours
Defocus Maps Estimation evaluation in terms of MSE and MAE						
MSE	**0.094**	0.172	0.098	0.205	0.374	0.350
MAE	**0.196**	0.311	0.239	0.358	0.509	0.457
Image deblurring comparison in terms of PSNR and SSIM						
PSNR	25.42	24.03	**26.08**	25.08	31.74	34.08
SSIM	**0.850**	0.761	0.823	0.797	0.812	0.908



(a). Input (b). Ground truth (c). Andres [12] (d). Lee [14] (e). Ma [4] (f). Ours

Figure 5.2 Defocus Maps Estimation comparison with [4], [10] and [11] on DED dataset.

where ε is a hyper-parameter set to $1e-3$. This hyper-parameter acts as a pseudo-Huber loss and smooths the errors smaller than ε.

Hence, the total loss L_{total} becomes a weighted sum of both losses as follow:

$$L_{total} = \lambda_1 L_{re} + \lambda_2 L_{smooth} + \lambda_3 L_{deblur} \tag{5.4}$$

Here, the weights are set as $\lambda_1 = 2$, $\lambda_2 = 0.5$, and $\lambda_3 = 0.01$.

5.2 Experimental Results

The quantitative comparison of our DME approach is presented in the upper half of Table 5.1.

In the case of the DED dataset[11], the calculation of errors after tests on this dataset could not be performed due to the unavailability of GT defocus maps and GT AiF images. However, visual results are depicted in Figure 5.2. Our self-supervised method manages

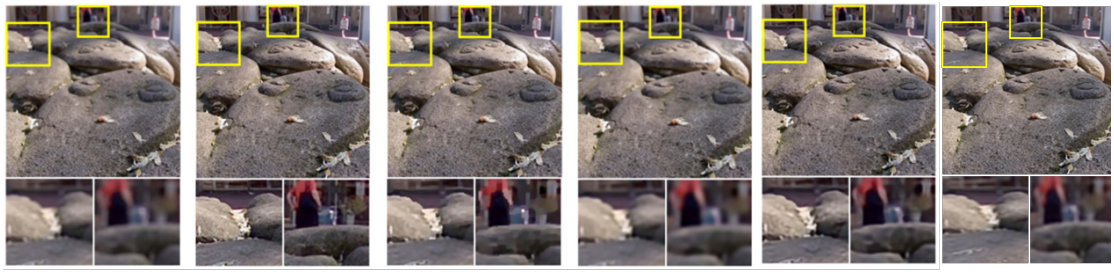


Figure 5.3 Image debluring comparison with [4], [10] and [11] on RTF while trained on DED dataset.

to recover a fair amount of features, the estimations being closer to the GT defocus maps, especially in comparison with [4].

For defocus deblurring, comparative results are provided on the RTF dataset in the lower half of Table 5.1. Evaluation metrics used include PSNR and SSIM. The methods in [4, 11] give better results but our method is more accurate than that of [10]. An example of defocus deblurring on RTF dataset is shown in Figure 5.3. The images from the RTF dataset are cropped and zoomed in small areas for better visualization and comparison. While [11] achieved impressive results in image deblurring, our method surpasses [10] in preserving fine details.

Chapter 6

Conclusions

In this thesis, we have developed a novel approach for depth estimation from a single defocused image using DNN, exploring defocus blur as a depth cue. Initially, we conducted a thorough analysis of DNNs in MDE, encompassing diverse deep architectures, pre-trained models, and loss functions. To improve the quality of depth maps, we studied several solutions for smoothing regularization used in DNN training. We compared their effectiveness in both supervised and self-supervised DNNs. Furthermore, we adapted the smoothing term in a self-supervised manner to function in a supervised approach. For experiments, we used as support a supervised DNN based on an Encoder-Decoder network. With our experiments, we show that the methods using regularization terms based on the L1 norm achieve the best accuracy. Instead of relying on common evaluation metrics the quality of edges has been measured by dedicated metrics consisting of the Kullback–Leibler Distance applied on the steering pyramid decomposition of depth maps. Our investigations in this regard were based on AiF images for training.

Subsequently, we harnessed the potential of defocus blur, an inherent phenomenon in the images taken with a lens camera, and we developed a multi-task DNN called 2HDED:NET, for depth estimation and image deblurring. This novel model is built upon the encoder-decoder architecture, by grafting a second decoder such to achieve in parallel, these two tasks. Comparing with previous methods relying on pipeline architectures with depth maps as intermediate results, 2HDED:NET demonstrated good performance producing accurate depth maps and recovering AiF images from single out-of-focus images. We extensively validated the proposed model on both indoor and outdoor benchmark datasets, showcasing its superiority over the baseline method [1].

There is a critical limitation when working with synthetic blur in the input images as it is the case for the existing benchmarks that include only AiF images and GT depth. This domain gap between real and synthetic datasets can lead to a significant decline in DNN performance. To address this issue, we introduced the iDFD dataset, containing naturally defocused images along with their corresponding AiF pairs and depth maps. Our experiments demonstrated the importance of training and testing DNNs on natural

scenes to close this domain gap. Additionally, we trained/tested various SoA depth estimation methods on the iDFD dataset and compared the results with the 2HDED:NET.

Initially, 2HDED:NET has been operated in a supervised manner, requiring GT depth maps and AiF images. In order to overcome the challenge of obtaining such a labor-intensive GT we transformed 2HDED:NET into a self-supervised training network by introducing a defocus simulation module that regenerates synthetically re-blurred images. This novel architecture enables the estimation of defocus maps in the absence of depth or defocus GT during the training.

6.1 Original contributions

In this thesis, we used defocus blur as a cue for improving the quality of depth estimation from single images. To this purpose, we propose a novel approach to perform depth estimation and image deblurring using a single out-of-focus image as an input. The main contributions of this thesis are:

- We proposed a novel deep learning-based architecture called 2HDED:NET with two parallel decoders that estimate depth and recover AiF images from a single out-of-focus image. The two-headed architecture distinguishes our network from existing methods that use pipeline processing. The task parallelization reduces the network complexity, all while maintaining good performances in depth estimation and deblurring comparable with SoA approaches. With extensive experiments, we show that the performance of the proposed method is competitive with the SoA approaches for both depth estimation and image deblurring.

- We proposed iDFD dataset which is a collation of naturally defocused indoor scenes that has the novelty of being supplemented by both depth and AiF GT. The tests on iDFD with the multi-task network 2HDED:NET, which simultaneously estimates the depth and deblurs the image, have proved that training a network on real rather than simulated data like NYU-Depth V2 synthetically defocused, is by far more effective. This dataset is useful also for training networks dedicated to image deblurring.

- We also proposed a novel method, that jointly estimates a defocus map and reconstructs an AiF image from a single defocused image. The guidance provided by Defocus Simulation Module enables the network to estimate the defocus map in a self-supervised way, image deblurring on the other hand is supervised. Experiments on various realistic and synthetic datasets show that our proposed self-supervised method obtains promising results comparing the SoA-supervised methods for both defocus map estimation and defocused image deblurring tasks, both quantitatively and qualitatively.

6.2 List of original publications

6.2.1 Journal Paper

Saqib Nazir[a], Lorenzo Vaquero[b], Manuel Mucientes [b], Víctor M. Brea [b], Daniela Coltuc [a]. *Depth Estimation and Image Restoration by DL From Defo*. In IEEE Transactions on Computational Imaging(TCI), volume. 9, pages. 607-619, 2023. ISSN: 1051-4651. DOI: 10.1109/TCI.2023.3288335.

[a] Research Center for Spatial Information (CEO SpaceTech), University POLITEHNICA of Bucharest (UPB), Romania.
[b] Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS) and Departamento de Electrónica e Computación, Universidade de Santiago de Compostela, Spain.

6.2.2 Conference Papers

Saqib Nazir[a], Cristian Damian [a], Daniela Coltuc [a]. *Self-supervised Defocus Map Estimation and Auxiliary Image Deblurring Given a Single Defocused Image*. In IEEE Digital Image Computing: Techniques and Applications (DICTA), volume. –, pages. –, 2023. ISSN: –. DOI: –submitted.

[a] Research Center for Spatial Information (CEO SpaceTech), University POLITEHNICA of Bucharest (UPB), Romania.

Saqib Nazir[a], Zhouyan Qiu[b], Daniela Coltuc [a], Joaquin Martinez-Sanchez [b], Pedro Arias [b]. *iDFD: A Dataset Annotated for Depth and Defocus.*. In Springier Cham, Scandinavian Conference on Image Analysis (SCIA), volume. 13885, pages. 6-19, 2023. ISSN: 978-3-031-31435-3. DOI: 1https://doi.org/10.1007/978-3-031-31435-3-5.

[a] Research Center for Spatial Information (CEO SpaceTech), University POLITEHNICA of Bucharest (UPB), Romania.
[b] Centro de Investigacion en Technologias, Energia y Procesos Industriales (CINTECX), Universidade de Vigo, Applied Geotechnology Group, Vigo, Spain.

Saqib Nazir[a], Lorenzo Vaquero[b], Manuel Mucientes [b], Víctor M. Brea [b], Daniela Coltuc [a]. *2HDED:Net for Joint Depth Estimation and Image Deblurring from a Single Out-of-Focus Image*. In IEEE International Conference on Image Processing (ICIP), pp. 2006-2010, Bordeaux (France), 2022. ISSN: 1051-4651. DOI: 10.1109/ICIP46576.2022.9897352.

[a] Research Center for Spatial Information (CEO SpaceTech), University POLITEHNICA of Bucharest (UPB), Romania.
[b] Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS) and Departamento de Electrónica e Computación, Universidade de Santiago de Compostela, Spain.

Saqib Nazir[a], Daniela Coltuc [a]. *Edge-preserving Smoothing Regularization for Monocular Depth Estimation*. In Proceedings of the 26th International Conference on Automation and Computing (ICAC), pp. 1–6, Portsmouth (United Kingdom), 2021. ISSN: 1051-4651. DOI: 10.23919/ICAC50006.2021.9594153.

[a] Research center for spatial information CEO SpaceTech, University POLITEHNICA of Bucharest (UPB), Romania.

6.3 Perspectives for further developments

In this thesis, we proposed 2HDED:NET, which benefits from out-of-focus blur to estimate depth maps. It shows promising results for the out-of-focus blur for DFD; however, the method is not able to estimate the depth maps in the presence of motion blur or camera shake. Hence, our future work will investigate depth estimation from camera shakes or motion blur. The AiFD head of the 2HDED:NET serves the purpose of out-of-focus deblurring; however, it is limited in handling the noise present in the defocus images. Our future work will also employ techniques that can address noise reduction during image deblurring.

On the one hand, defocus blur is an important cue for depth estimation but on the other hand, it degrades the image and affects the quality of the image. With our self-supervised 2HDED:NET we managed to generate the defocus maps without GT depth/defocus information but we carry our interest towards image restoration in a complete self-supervised manner. Because there are many cases in the practical world, where a sharp pair of blurred images is not available, hence, future developments could also include the development of self-supervised training for both the AiF and defcous map outputs. A promising option would be to use this approach as the first stage of a training process that is further refined by a supervised fine-tuning stage. Further developments also focus on progressing toward a self-supervised approach for depth estimation by adding supplementary information such to remove defocus uncertainty.

The iDFD dataset proposed in this thesis is designed for depth estimation and image deblurring. While this dataset is comprehensive and can serve as a valuable resource for training DNNs, it currently encompasses scenes exclusively from indoor environments. However, in real-world scenarios, we encounter challenging scenes in both indoor and outdoor environments. Consequently, potential future research in this thesis involves extending the iDFD dataset to include outdoor environments. In order to capture the GT depth from the outdoor scenes a specialized sensor such as LiDAR is required. This expansion would enhance the dataset's applicability and foster the development of DNN models capable of tackling diverse and complex real-world scenes. This step is crucial to provide future studies in the field of DFD and image deblurring with quantitative benchmarks based on more realistic data.

References

[1] Anwar, S., Hayder, Z., and Porikli, F. (2021). Deblur and deep depth from single defocus image. *Machine vision and applications*, 32(1):1–13.

[2] Carvalho, M., Le Saux, B., Trouvé-Peloux, P., Almansa, A., and Champagnat, F. (2018). 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) Workshops*, pages 0–0.

[3] Charbonnier, P., Blanc-Feraud, L., Aubert, G., and Barlaud, M. (1994). Two deterministic half-quadratic regularization algorithms for computed imaging. In *Proceedings of 1st International Conference on Image Processing*, volume 2, pages 168–172. IEEE.

[4] D'Andrès, L., Salvador, J., Kochale, A., and Süsstrunk, S. (2016). Non-parametric blur map regression for depth of field extension. *IEEE Transactions on Image Processing*, 25(4):1660–1673.

[5] Eigen, D., Puhrsch, C., and Fergus, R. (2014). Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27.

[6] Godard, C., Mac Aodha, O., and Brostow, G. J. (2017). Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 270–279.

[7] Godard, C., Mac Aodha, O., Firman, M., and Brostow, G. J. (2019). Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3828–3838.

[8] Gur, S. and Wolf, L. (2019). Single image depth estimation trained via depth from defocus cues. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7683–7692.

[9] Hu, J., Ozay, M., Zhang, Y., and Okatani, T. (2019). Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries. In *2019 IEEE winter conference on applications of computer vision (WACV)*, pages 1043–1051. IEEE.

[10] Lee, J., Son, H., Rim, J., Cho, S., and Lee, S. (2021). Iterative filter adaptive network for single image defocus deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2034–2042.

[11] Ma, H., Liu, S., Liao, Q., Zhang, J., and Xue, J.-H. (2021). Defocus image deblurring network with defocus map estimation as auxiliary task. *IEEE Transactions on Image Processing*, 31:216–226.

[12] Nazir, S. and Coltuc, D. (2021). Edge-preserving smoothing regularization for monocular depth estimation. In *2021 26th International Conference on Automation and Computing (ICAC)*, pages 1–6. IEEE.

[13] Nazir, S., Qiu, Z., Coltuc, D., Martínez-Sánchez, J., and Arias, P. (2023a). idfd: A dataset annotated for depth and defocus. In *Scandinavian Conference on Image Analysis*, pages 67–83. Springer.

[14] Nazir, S., Vaquero, L., Mucientes, M., Brea, V. M., and Coltuc, D. (2022). 2hded: Net for joint depth estimation and image deblurring from a single out-of-focus image. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 2006–2010. IEEE.

[15] Nazir, S., Vaquero, L., Mucientes, M., Brea, V. M., and Coltuc, D. (2023b). Depth estimation and image restoration by deep learning from defocused images. *arXiv preprint arXiv:2302.10730*.

[16] Qiu, Z., Martínez-Sánchez, J., Brea, V. M., López, P., and Arias, P. (2022). Low-cost mobile mapping system solution for traffic sign segmentation using azure kinect. *International Journal of Applied Earth Observation and Geoinformation*, 112:102895.

[17] Si, H., Zhao, B., Wang, D., Gao, Y., Chen, M., Wang, Z., and Li, X. (2023). Fully self-supervised depth estimation from defocus clue. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9140–9149.

[18] Ummenhofer, B., Zhou, H., Uhrig, J., Mayer, N., Ilg, E., Dosovitskiy, A., and Brox, T. (2017). Demon: Depth and motion network for learning monocular stereo. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5038–5047.

[19] Wang, Z. and Simoncelli, E. P. (2005). Reduced-reference image quality assessment using a wavelet-domain natural image statistic model. In *Human vision and electronic imaging X*, volume 5666, pages 149–159. SPIE.

[20] Xian, K., Zhang, J., Wang, O., Mai, L., Lin, Z., and Cao, Z. (2020). Structure-guided ranking loss for single image depth prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 611–620.

[21] Zhang, K., Zuo, W., Chen, Y., Meng, D., and Zhang, L. (2017a). Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE transactions on image processing*, 26(7):3142–3155.

[22] Zhang, K., Zuo, W., Gu, S., and Zhang, L. (2017b). Learning deep cnn denoiser prior for image restoration. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3929–3938.