Depth Estimation and Image Restoration by Deep Learning From Defocused Images

Saqib Nazir, *Graduate Student Member, IEEE*, Lorenzo Vaquero, Manuel Mucientes, Víctor M. Brea, and Daniela Coltuc

***Abstract*—Monocular depth estimation and image deblurring are two fundamental tasks in computer vision, given their crucial role in understanding 3D scenes. Performing any of them by relying on a single image is an ill-posed problem. The recent advances in the field of Deep Convolutional Neural Networks (DNNs) have revolutionized many tasks in computer vision, including depth estimation and image deblurring. When it comes to using defocused images, the depth estimation and the recovery of the All-in-Focus (Aif) image become related problems due to defocus physics. Despite this, most of the existing models treat them separately. There are, however, recent models that solve these problems simultaneously by concatenating two networks in a sequence to first estimate the depth or defocus map and then reconstruct the focused image based on it. We propose a DNN that solves the depth estimation and image deblurring in parallel. Our Two-headed Depth Estimation and Deblurring Network (2HDED:NET) extends a conventional Depth from Defocus (DFD) networks with a deblurring branch that shares the same encoder as the depth branch. The proposed method has been successfully tested on two benchmarks, one for indoor and the other for outdoor scenes: NYU-v2 and Make3D. Extensive experiments with 2HDED:NET on these benchmarks have demonstrated superior or close performances to those of the state-of-the-art models for depth estimation and image deblurring.**

***Index Terms*—Depth from defocus, image deblurring, deep learning.**

I. Introduction

DEPTH estimation from a single image is a key problem in computer vision, where it spans a lot of applications. Robotics, augmented reality, human-computer interaction or computational photography, to give only several examples, benefit from depth estimation. With the recent advancements in 3D computer vision and the newly emerging tasks like semantic segmentation or 3D object detection, depth estimation has become even more important.

The depth can be measured by specialized devices or can be inferred from images and videos. For outdoor scenes, LIDAR or stereo systems are typically used to measure the depth of the scene. For indoor scenes, Time of Flight (ToF) cameras like RGBD Kinect from Microsoft, is used to capture depth information in addition to the RGB images. However, the applicability of these devices is limited. ToF cameras are not working properly in the outdoors, being limited to 30 m at best, while the LIDAR may produce poor-quality depth maps because of infrared interference. These physical limitations, the sparse nature of the measurements, and the cost of the devices have fostered the research in the direction of obtaining depth from images or videos taken with commercial cameras. Here, although the performance of depth estimation methods is steadily increasing, there are still major problems related to the accuracy and resolution of the estimated depth maps.

Image deblurring is a classical problem in low-level computer vision, and a prepossessing step in numerous applications such as face detection, classification, object recognition, or misfocus correction. Object motion, camera shake, or out-of-focus are common causes for the blur appearing in the images taken with a camera. The goal of image deblurring is to recover an AiF image with all the details and sharp edges from its defocused counterpart.

The main objective of the network proposed in this article is to estimate the depth and remove blur from a single out-of-focus image. Fig. 1 shows an example of Depth from Defocus (DFD) and image deblurring. In this example, the network estimates a dense depth map and reconstructs the AiF image from a defocused image.

Most of the DNNs dedicated to depth estimation work on AiF images [1], [2], [3]. The exploitable information in such images is limited to the scene geometry, which explains the lower performances compared to LIDAR or ToF cameras. The defocus blur is a complementary cue that can help to improve the depth accuracy.

DFD has been widely investigated in the past [4]. The first DFD methods were focused on the depth related to the blur amount and as a result, they suffered from insensitivity in the Depth of Field (DoF) region and uncertainty regarding the object

Manuscript received 30 May 2022; revised 13 October 2022, 7 March 2023, and 2 May 2023; accepted 27 May 2023. Date of publication 21 June 2023; date of current version 29 June 2023. The work of Lorenzo Vaquero was supported by the Spanish Ministerio de Universidades under the FPU National Plan FPU18/03174. The work of Daniela Coltuc was supported by UEFISCDI Romania under Grant 31/01.01.2021 PN III, 3.6 Support. This work was supported in part by the European Union's Horizon 2020 Research and Innovation Programme under the Marie Skłodowska-Curie under Grant 860370, in part by the Spanish Ministerio de Ciencia e Innovación under Grant PID2020-112623GB-I00, and in part by the Galician Consellería de Cultura, Educación e Universidade co-funded by the European Regional Development Fund (ERDF), under Grants ED431C 2018/29, ED431C 2021/048, and ED431G 2019/04. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Jinli Suo. *(Corresponding author: Saqib Nazir.)*

Saqib Nazir and Daniela Coltuc are with the CEOSpaceTech, University POLITEHNICA of Bucharest (UPB), 060042 Bucharest, Romania (e-mail: saqib.nazir@upb.ro; daniela.coltuc@upb.ro).

Lorenzo Vaquero, Manuel Mucientes, and Víctor M. Brea are with the Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), University of Santiago de Compostela (USC), 15705 Santiago de Compostela, Spain (e-mail: lorenzo.vaquero.otal@usc.es; manuel.mucientes@usc.es; victor.brea@usc.es).

Digital Object Identifier 10.1109/TCI.2023.3288335

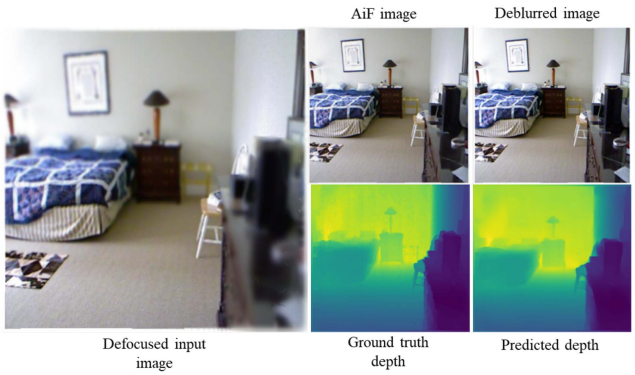

Fig. 1. Example of image deblurring and depth prediction using 2HDED:NET in a scene from the NYU-v2 dataset. AiF is an all-in-focus image that serves as ground truth for deblurring. The AiF and ground-truth depth are captured by an RGBD Kinect camera.

position with respect to the in-focus plane. The use of coded apertures [5], [6], dual images or focal stacks [7], [8], [9] has alleviated such problems.

In many applications, including DFD [10], [11], [12], [13], DNN models outperform the classical methods, due to the ability of learning more complex features. The features learned from defocused images combine both the scene geometry and the blur for more accurate depth estimation. In the last years, a series of DNNs has been proposed for image deblurring as well [10], [14], [15], [16], [17].

Although the depth and defocus blur are closely related, the deblurring and depth estimation have been generally, treated as separate problems by deep learning. There are however some rare exceptions like the method of Anwar et al. [10], which concatenates two networks to first estimate the depth map and then, based on this depth map, restores a focused image by pixel-wise non-blind deconvolution. More precisely, in [10] a fully convolutional neural network with 13 layers provides a pixel level feature map, then a patch pooling layer turns the patches around predefined key points into fixed size feature map, which are further propagated through a shallow fully connected network to estimate a dense depth map. The deblurring is done by deconvolution with kernels calculated for every pixel of the RGB image, by using the estimated depth.

It is known that architectures with independent and task-dedicated branches and their loss terms combined decrease overfitting in the training phase, and permit to execute any of the tasks in the inference time. In this line, we propose a DNN that solves the problem of DFD and image deblurring in parallel. The proposed two-headed network, called 2HDED:NET, estimates the depth and deblurs the image in a balanced way by giving the same importance to both tasks. The network consists of three modules: i) an encoder for multi-level feature extraction from the defocused image, ii) a depth estimation decoder (DED) for the DFD, iii) an AiF decoder (AifD) for image deblurring (Fig. 2). The heads interact with each other during training, allowing the encoder to learn semantically rich features that are well-suited for both tasks.

Unlike Anwar et al. in [10], where the deblurring depends on the intermediate result of depth estimation, our 2HDED:NET generates an AiF image, which is not any more constrained by depth accuracy. Separating the deblurring and depth estimation branches also makes AifD self-sufficient and better able to perform the deblurring task without relying on an estimated depth map.

2HDED:NET is a typical Multitask Learning (MTL) neural network with hard sharing of parameters. The encoder layers are shared by both depth estimation and deblurring tasks while the two decoders remain task-specific. Comparing to the single task networks, the MTL networks benefit from a series of advantages: an augmented training set, relevant feature learning by attention focusing, easier learning of features from less complex models, reduced risk of overfitting, and better generalization to new tasks [21]. The foundations of MTL by hard parameter sharing had been laid by Caruana in 1997 [18], and two recent surveys of MTL can be found in [19], [20]. The MTL technique has been used successfully in computer vision applications as well as in other areas such as natural language processing and drug discovery. Two recent applications closely related to our application are addressed in [21], [22], where the depth map and semantic segmentation are learned by MTL.

The architecture of 2HDED:NET is straightforward, simple, and easy to train. With its double functionality – depth estimation and deblurring – 2HDED:NET emulates a Kinect-type camera on a commercial camera with limited DoF. A special feature of 2HDED:NET is that after training, the depth estimation head is no longer necessary to recover a sharp image and vice versa.

We define a hybrid loss function to train 2HDED:NET. It embeds specific cost functions for depth and deblurring like $L1$ norm and Charbonier loss [11], [23], as well as specific regularizations like gradient-based smoothing [24] and maximization of Structural Similarity Index Measure (SSIM).

We run extensive experiments on the NYU-v2 and Maked3D benchmarks in order to evaluate the performance of the 2HDED:NET and to compare with the state-of-the-art methods for depth estimation and image deblurring. In most cases, 2HDED:NET generates better results. For training, 2HDED:NET uses two types of ground truth, the depth and the AiF images in the benchmarks. As input, synthetically defocused images are generated using the thin lens model. Hence, the prior information is consistent with that of any other network for depth estimation. The supplement includes the mathematical model of the defocus blur, which is proven effective through our results.

The main scientific contributions of our work are:

- A parallel architecture, namely 2HDED:NET, that enables the recovery of AiF images and the generation of depth maps from a single defocused image.
- The architecture has the merit to achieve a balanced generation of both depth maps and AiF images while assigning equal significance to both tasks.
- A hybrid loss function that combines losses and regularizations from both depth estimation and deblurring and enforces the encoder to learn much richer semantic features.
- Extensive experimental results on NYU-v2 and Make3D datasets enriched with synthetic defocused images, confirm the effectiveness of our approach.

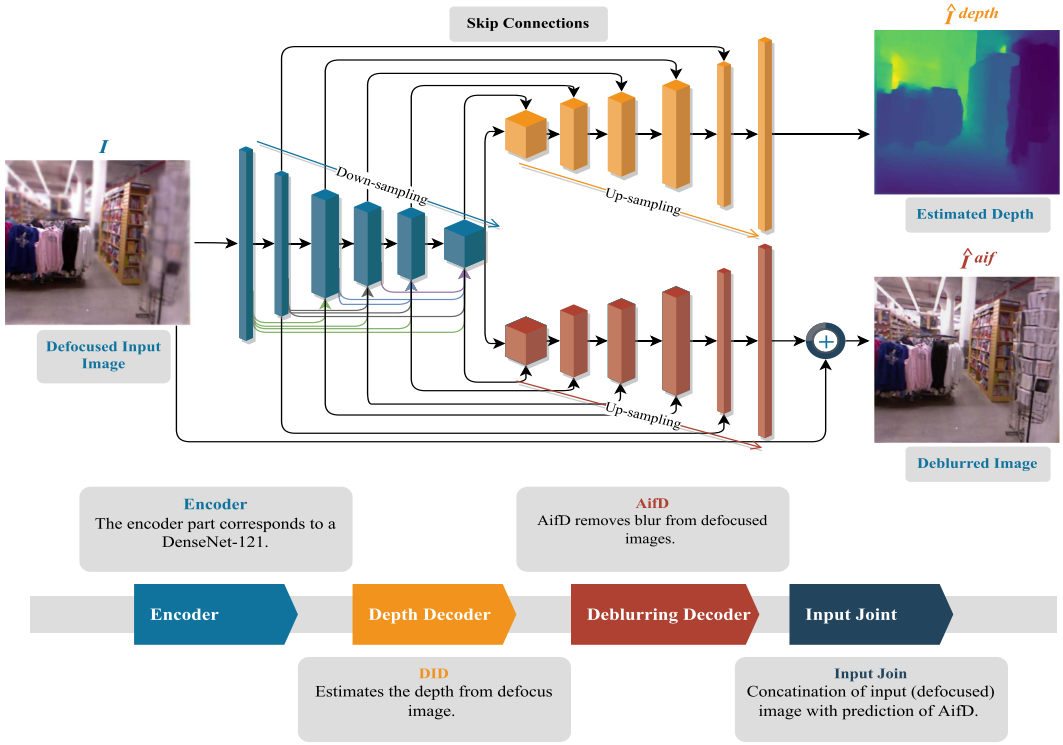

Fig. 2. 2HDED:NET architecture consists of one encoder and two decoders that work in parallel. The upper Head estimates the depth map and the lower one the AiF image. The network is fed in with defocused RGB images.

The remainder of this article is organized as follows. In Section II, we provide an overview of the related work. In Section III, we present our methodology. The experimental setup and results are reported in Sections IV and V, and, finally, in Section VI, we present conclusions and future work.

II. Related Work

This section briefly reviews the state-of-the-art DNN-based solutions for DFD and deblurring.

A. Single Image Depth Estimation

The success of DNN models in various fields of computer vision, such as image segmentation and classification, has prompted the scientific community to consider using DNNs for depth estimation as well. Saxena et al. [2] presented one of the first solutions for monocular depth estimation with deep learning methods. They estimate the depth with a multi-scale architecture and the Markov Random Field (MRF). Eigen et al. [3] presented one of the most successful works by developing a multi-scale architecture to extract information from a scene at global and local levels in order to estimate the depth map. Laina et al. [25] proposed an encoder-decoder network with a fast up-projection block. Cao et al. [26] relied on Conditional Random Fields (CRF) to improve the accuracy of the depth maps. GANs have also been utilized for depth prediction. Jung et al. in [27] and Carvalho et al. in [28] implemented an adversarial loss for depth prediction. Most of the evoked networks solve the problem of monocular depth estimation by using in-focus images as input and ignoring the defocus blur, which is however an important cue in depth estimation.

B. Depth From Defocus

Defocus blur occurs when images are captured with limited DoF. All cameras have a limited DoF, which is controlled by the camera aperture diameter. Although the blur exists also in this range, it is not perceived by the human eye.

To estimate depth from a single defocused image, Carvalho et al. [11] built on a dense network DenseNet-121 with skip connections that improved the state-of-the-art results of the time. To handle the DFD problem, Gur et al. [29] designed a convolution layer based on the Point Spread Function (PSF) to train an unsupervised network. Anwar et al. [10] trained a fully connected cascaded deep neural network inspired by the VGG-16 model on dense overlapping fields to estimate depth from a single defocused image. In [12], Fu et al. proposed a multi-scale network structure to obtain high-resolution depth maps using spacing-increasing discretization and a simple regression loss.

C. Image Deblurring

Blind image deblurring has been always a difficult problem. Since the advent of DNNs, several models were designed to reduce the blur from a single image. The first ones directly remove the blur, such as Nah et al. [30], who used a multiscale loss function to train their model. Tao et al. [31] improved their work by using joint network parameters at different scales. Kupyn

et al. proposed DeblurGAN to reconstruct AiF images from defocused images by using an adversarial loss function [32].

The strategy adopted in more recent papers is to first estimate a defocus/depth map and later use this information for image deblurring. Thus, Lee et al. [33] introduced a deep architecture along with a domain-matching approach to estimate the defocus map of an image, and also presented a large dataset for training DNNs. Very recently, by estimating defocus maps, works like [10], [14] use the amount of blur per pixel to reconstruct the entire deblurred image.

Zhang Kai et al. proposed two general methods for image restoration, with deblurring as a particular case in [34], [35]. In [34], it is shown that by separating the fidelity from the regularization term in the energy function, the optimization problem can be solved by plugging a denoising neural network in a Half Quadrating Splitting framework. The method is tested for denoising, super-resolution and deblurring. In [35], the authors proposed a convolutional neural network for blind Gaussian denoising. The network removes the latent clean image and estimates the residual Gaussian noise with unknown level. By observing that the image degradation model for Gaussian denoising can be converted to other restoration problems, the authors successfully apply it to image super-resolution and JPEG deblocking.

The priors on image model play an important role in image restoration by optimization. Zha et al. proposed in [36] a low-rank and deep image model with three complementary priors: internal and external, shallow and deep, and non-local and local priors. The model is successfully tested on image deblurring, restoration after compressive sensing, and JPEG deblocking. The sole non-local self-similarity prior is used by Zha et al. in [37] for image restoration by using the Expectation Maximization algorithm with image deblurring, denoising, and deblocking as applications.

D. Joint DFD and Image Deblurring

The survey of the literature has revealed only two DNN models that addressed both depth estimation and blurring or deblurring process. Gur et al. proposed a network to estimate the depth from a single defocused image in [29]. Unlike the supervised learning networks, they adopted a self-supervised learning approach with a loss function based on the difference between the defocused input image and a defocused image estimated by a second network. This second network implements the blur model and creates a synthetically defocused image by using the estimated depth.

The closest approach to our architecture is the model proposed by Anwar et al. in [10]. They train a cascade of two smaller networks to estimate a depth map, which is then used to compute kernels for restoring the AiF image by pixel-wise non-blind deconvolution.

III. 2HDED:NET Architecture

Fig. 2 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 Fig. 2, 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 ground truth depth as well as the AiF images for training.

TABLE I
Size of Output Features and Input/Output Channels of Each Layer of 2HDED:NET

Layer	*Output size*	*Input/C*	*Output/C*
Conv1	128×128	3	64
Conv2	64×64	64	128
Conv3	32×32	128	256
Conv4	16×16	256	512
Conv5	8×8	512	1024
Dconv5	8×8	1024	1024
Dconv4	16×16	1024	512
Dconv3	32×32	512	256
Dconv2	64×64	256	128
Dconv1	128×128	128	64
Pred-depth	256×256	64	1
Pred-deblurring	256×256	64	3

A. Encoder

For the encoder network, we use the DenseNet-121 [38]. As its name suggests, DenseNet consists of densely connected layers. The main feature of DenseNet-121 is that this network reuses the features of each layer by concatenating them with the features of the next layer, rather than simply aggregating them like ResNet50. The goal of the concatenation is to use the features obtained in the previous layers in the deeper layers. This is referred to as "feature reusability". DenseNets can learn mappings with fewer parameters than a typical CNN since there are no redundant maps to learn. Similar to [11], we replace the max-pooling layer with a 4×4 convolutional layer to reduce resolution while increasing the number of the feature channel maps. We use skip connections between the encoder and decoder parts to simplify learning. The skip connections prevent the problem of the gradient disappearing since the subsequent layers focus on solving residuals rather than completely new representations. The encoder helps to obtain multi-resolution features from the input image, which are useful for the two tasks that 2HDED:NET performs. Further information about the encoder's output size, input, and output channels can be seen in Table I.

B. Depth Estimation Head

The Depth Estimation Decoder (DED) is inspired by [11]. It consists of 5 decoding layers, each with 4×4 convolution that increases the resolution of the feature map, 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. Table I shows how decoder layers upsample the input using transpose convolutions. The output of DED is one channel depth map.

C. Deblurring Head

We refer to the deblurring decoder as to AiF decoder (AifD). Unlike DED, the output of AifD is a three-channel RGB image. We use an input joint layer to aggregate the defocused input image with the output of AifD as in [14], [39] 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 our solution is that once 2HDED:NET is fully trained, we are still able to perform a task when the other head is removed, e.g. we can perform DFD without the AifD head and vice versa.

D. Loss Functions

The training of the 2HDED:NET is supervised simultaneously by ground truth depth maps and AiF images. To consider this dual information, we propose a loss function 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.

1) Depth Loss: Most of the deep learning 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 [11].

As the loss function for depth estimation, we resort to $L1$ norm, known for the ability to estimate sparse solutions as it is the case for depth maps [11], [40]:

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

where $\widehat{I}^{depth}$ is the estimated depth, I^{depth} the ground truth, 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 [24], [29], [41], [42]. In the case of our network, we improve the depth accuracy by combining $L1$ norm with the smoothing term commonly used in supervised learning and defined as [24]:

$$L_{grad} = \frac{1}{n}\sum_{i} |\Delta_x R_i| + |\Delta_y R_i| \tag{2}$$

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):

$$L_{depth} = L_1^{Depth} + \mu L_{grad} \tag{3}$$

where μ is a weighting coefficient set to 0.001.

2) Deblurring Loss: 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 [43], [44].

To train 2HDED:NET, we test $L1$ norm and Charbonnier loss function [45], which is the smoothed version of $L1$. Charbonnier loss is calculated as a squared error between the estimated deblurred image $\widehat{I}^{aif}$ and the ground truth AiF image I^{aif}:

$$L_{charb} = \frac{1}{n}\sum_{i=1}^{W}\sum_{j=1}^{H}\sqrt{(\widehat{I}_{i,j}^{aif} - I_{i,j}^{aif})^2 + \epsilon^2} \tag{4}$$

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.

In a series of papers [23], [39], [46], the loss function defined either as Charbonnier or $L1$ norm, is improved by requiring a high SSIM. This results in adding the regularization term:

$$L_{SSIM} = 1 - SSIM\left(\widehat{I}_{i,j}^{aif}, I_{i,j}^{aif}\right) \tag{5}$$

which makes the complete deblurring loss function to be:

$$L_{deblur} = L_{charb} + \Psi L_{SSIM} \tag{6}$$

where Ψ is a weight set to 4.

3) 2HDED Loss Function: With the depth and deblurring losses defined as in (3) and (6), we define the following total loss for 2HDED:NET training:

$$L_{2HDED} = L_{depth} + \lambda L_{deblur} \tag{7}$$

In our experiments, we tested several versions of L_{2HDED}: with L_{depth} including or not L_{smooth}, with L_{deblur} being either $L1$ norm or L_{charb}, and with or without $SSIM$ loss. We noticed during the experiments that the model performance is very sensitive to the weighting value, which is why we paid close attention to the choice of λ. Starting from the idea that both tasks should be given the same importance, we evaluated the depth and deblurring losses separately during the training, and we settled λ in a way that they have an approximately similar contribution to the total loss. Then we fine-tuned λ by performing a grid search and we found that $\lambda = 0.01$ is suitable for all versions of L_{HDED}.

E. Accuracy Measures for DFD and Image Deblurring

To evaluate the accuracy of the estimated depth maps and deblurred images, we use accuracy measures that have been widely reported in previous studies.

For the depth estimation, we compute the root mean square error (RMSE), relative absolute error (Abs. Rel.), and thresholded accuracy δ as follows:

1) $RMSE = \sqrt{\frac{1}{n}\sum_{i=0}^{n}(\widehat{I}_t^{depth} - I^{depth})^2}$
2) $Abs.Rel. = \frac{1}{n}\sum_{i=0}^{n}\frac{|\widehat{I}_t^{depth} - I^{depth}|}{I^{depth}}$
3) Thresholded accuracy (δ) is the percentage of pixels such that: $\max(\frac{\widehat{I}_t^{depth}}{I^{depth}}, \frac{I^{depth}}{\widehat{I}_t^{depth}}) = \delta < threshold$

To evaluate the deblurring, we resort to two well-known metrics commonly used to measure the quality of images: Peak Signal to Noise Ratio (PSNR) and SSIM.

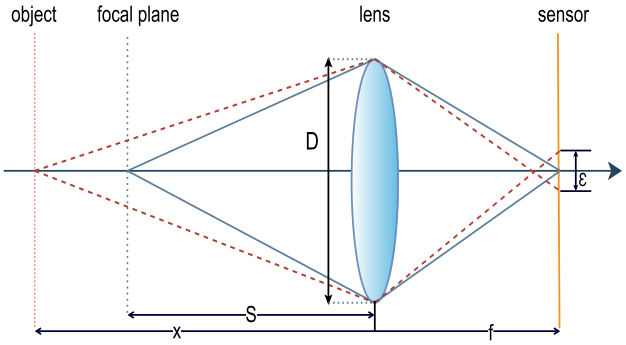

Fig. 3. The thin lens model: the COC diameter ε depends on the distance x of the object to the lens.

IV. Datasets

2HDED:NET is trained with two types of ground truth, depth maps, and AiF images. Since until recently, DFD and image deblurring have been considered separately, the existing solutions were developed around datasets dedicated to one of these applications. The lack of datasets including defocused images, corresponding depth maps, and AiF images, determined us to work on datasets for depth applications consisting of AiF images and depth ground truth and to generate defocused images by blurring the AiF images. The synthetically defocused images have been used by many recent works [10], [11], [13] dedicated either to depth inference or image restoration. Thus, our choice has been the NYU-Depth V2 dataset containing indoor scenes, and the Make3D dataset with outdoor scenes. The depth range of the datasets depends on the type of sensor used to capture the depth as well as the collection method. The depth range of NYU images is 0.7 to 10 m and of Make3D, 0 to 80 m.

The NYU dataset comprises 230,000 pairs of RGB indoor images and their corresponding depth maps. In order to speed up the experiment, the training of 2HDED:NET has been run with a smaller dataset. We used the same split as [10], [11], [47] i.e., 795 images for training and 654 for testing. The original size of the images captured by Microsoft Kinect is 640×480 pixels, but they were reduced to 561×427 pixels in our experiments.

The Make3D dataset consists of 534 RGB images and depth maps representing outdoor scenes. To train the 2HDED:NET under the same conditions as in [10], we split the dataset similarly, i.e., into 400 images for training and 134 images for testing.

To avoid overfitting, the training set has been increased by data augmentation. We adopted the data augmentation procedure addressed in [11]. Since we use defocus blur as a cue, we do not apply any data augmentation process that can affect the blur information. In the first step, all the images are centered scaled. For random flips, each individual sample is flipped horizontally by 50%.

A. Defocus Blur Simulation

To generate the realistic physical blur in the RGB images we adopt the procedure used by the authors in [48] to generate the SYNDOF dataset. To defocus an image, they start from the thin lens model [49], commonly used in computer vision (Fig. 3). We used the same parameter values as [48] such as aperture size of 4.48 cm and focal length set to 0.07 m. In Fig. 3, x is the distance to the object, f is the distance from the lens to the image sensor, D is the diameter of the aperture, S is the distance to the in-focus plane, and ε is the diameter of the circle of confusion (COC) calculated as:

$$\varepsilon = \alpha \frac{|x - S|}{x}, \text{where } \alpha = \frac{f}{S} D \tag{8}$$

To generate blur in the AiF image, we apply Gaussian filters with a kernel with standard deviation $\rho = \varepsilon/4$. Similar to [48], ε is calculated based on the per-pixel depth values. As a result, we have defocused images with corresponding depth maps and AiF images.

V. Results

For the experimental results, we divided our analysis into the following sections:

- Depth estimation and image deblurring results with various loss functions. We tested simple solutions like L_{charb} for deblurring and L_1 for depth and improved our results gradually, by adding regularizations consisting of SSIM for deblurring and smoothing for depth.
- Results with two heads and one head ablated to see the effectiveness of the two-head architecture.
- Finally, we compare our results to the state-of-the-art results for depth or image deblurring, obtained on the NYU-v2 and Make3D benchmarks.

Our network is implemented using the PyTorch package in Python environment The entire training session takes approximately 9 hours on an NVIDIA Quadro GV100 GPU with 32 GB memory. We trained 2HDED:NET for 500 epochs with a batch size of 4 images. We use Stochastic Gradient Descent (SGD) optimizer with an initial learning rate of 0.0002. The initial learning rate is reduced 10 times after the first 300 epochs, this allows for large weight changes at the beginning of the learning process and small changes towards the end of the learning process. As for the total number of network parameters, our network is much lighter than [10]. Specifically, our 2HDED:NET comprises a total of 41 M parameters, whereas [10] employs a network with 138 M parameters, three times higher than our model.

A. Effect of Loss Functions

In this subsection, we perform a set of experiments consisting of training the 2HDED:NET with the various loss functions described in Section III-D. To select among the multiple options existing for our combined depth and deblur loss function, we adopt a simple to complex approach. In the first step, we use a photometric error for deblurring and $L1$ norm for depth accuracy. For deblurring, we test with $L1$ and Charbonnier loss functions, the latter being a smoothed version of $L1$. To ensure an equal contribution of the two components, the deblurring error is weighted by $\lambda = 0.01$. We maintain the principle of equal contribution over the entire experiment.

Table II presents results obtained on the NYU dataset. The depth estimation has a good accuracy even for this simple loss

TABLE II
RESULTS ON NYU-V2 DATASET WITH DIFFERENT LOSS FUNCTIONS FOR DEPTH ESTIMATION AND IMAGE DEBLURRING

Loss Function	Depth Estimation					Image Deblurring	
	RMSE ↓	*Abs. rel* ↓	$\delta(1)$ ↑	$\delta(2)$ ↑	$\delta(3)$ ↑	*PSNR* ↑	*SSIM* ↑
$L_1^{Depth} + \lambda L_{charb}^{Deblur}$	0.285	0.035	0.820	0.880	0.970	33.55	0.983
$L_1^{Depth} + \lambda L_1^{Deblur}$	0.292	0.068	0.799	0.819	0.891	30.38	0.90
$(L_1^{Depth} + \mu L_{grad}) + \lambda L_{charb}^{Deblur}$	0.244	0.029	0.901	0.971	0.989	32.27	0.918
$L_1^{Depth} + \lambda(L_{charb}^{Deblur} + \Psi(1 - SSIM))$	0.282	0.031	0.833	0.895	0.901	33.85	0.981
$(L_1^{Depth} + \mu L_{grad}) + \lambda(L_{charb}^{Deblur} + \Psi(1 - SSIM))$	**0.241**	**0.025**	**0.914**	**0.979**	**0.995**	**34.84**	**0.989**

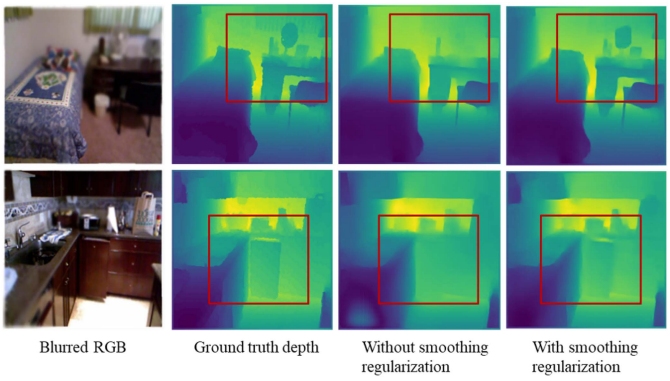

Fig. 4. Results with and without the smoothing regularization for depth estimation on NYU-v2 dataset. The deblurring loss is L_{charb}. The rectangular crops illustrate areas where new details emerge when using smoothing regularization.

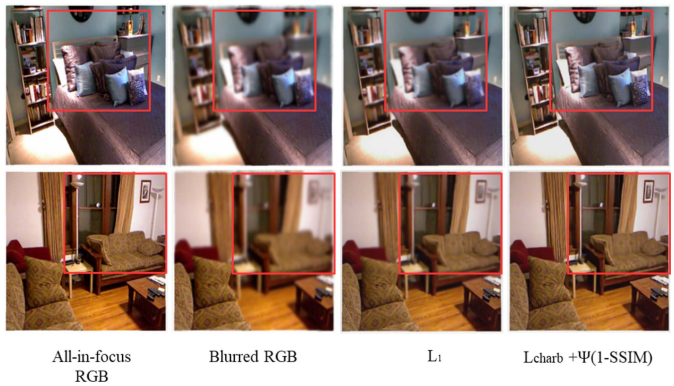

Fig. 5. Deblurring results with L_1 and $L_{charb} + \Psi(1 - SSIM)$ default loss on NYU-v2 dataset. The depth loss is L_1.

function. The RMSE is under 0.3 and there are no significant differences when the photometric error for deblurring switches from Charbonnier to $L1$, where the accuracy is only slightly worse. This is not the case for deblurring, where the use of Charbonnier improves the PSNR by 3 dB comparing with $L1$, providing on average, a quality of 33.554 dB for the test set. Therefore, we choose Charbonnier for the subsequent experiments.

In the second step, we alternately improve on deblur and depth losses by adding an SSIM-based term to L_{charb} and smoothing regularization to $L1$. These additional terms are weighted by $\Psi = 4$ and $\mu = 0.001$, respectively. The smoothing regularization improves the RMSE depth accuracy from 0.285 to 0.244 on average but it worsens the deblurring results by more than 1 dB. This apparently small difference in RMSE can impact significantly on the quality of depth maps as it can be seen from the example in Fig. 4, where new details are emerging when the smoothing regularization is added.

The introduction of SSIM term brings benefits to both depth estimation and deblurring. The RMSE decreases to 0.282 and the PSNR becomes higher by 0.3 dB on average. The deblurring results on simple L_1 loss and the default loss, which is $L_{charb} + \Psi(1 - SSIM)$, can be seen in Fig. 5.

Finally, we combine the photometric errors and the two regularizations – SSIM and smoothing – in a unique loss function. The network trained with this loss function achieves the best results both in depth accuracy and deblurring. The RMSE of depth touches the lowest level of 0.241 and the PSNR of the deblurred images is almost 35 dB. Some examples are depicted in Figs. 6 to 8 and commented in the next subsection.

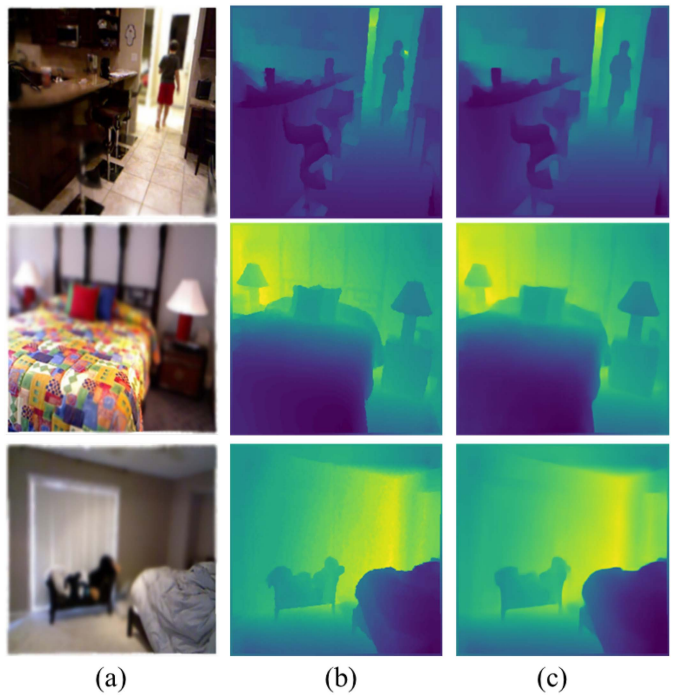

Fig. 6. 2HDED:NET results for depth estimation on NYU-v2 dataset using the default loss L_{2HDED}: (a) RGB defocused image (b) Depth ground truth (c) Estimated Depth.

B. *Effect of Head Ablation*

The 2HDED:NET is trained by using two kinds of ground truth values, the AiF image and the depth map. One is ingested by the deblurring head, the other one by the depth head. They

TABLE III
EFFECT OF ABLATING ONE HEAD. RESULTS ON NYU-V2 DATASET

Image Debluring					
	PSNR↑		*SSIM*↑		
With both heads	**34.849**		**0.989**		
Without depth head	31.941		0.919		
Gain	2.899		0.07		
Depth Estimation					
	RMSE ↓	*Abs. rel* ↓	$\delta(1)$ ↑	$\delta(2)$ ↑	$\delta(3)$ ↑
With both heads	**0.24**	**0.025**	**0.91**	**0.97**	**0.99**
Without deblurring head	0.29	0.075	0.84	0.89	0.94
Gain	0.05	0.05	0.07	0.08	0.05

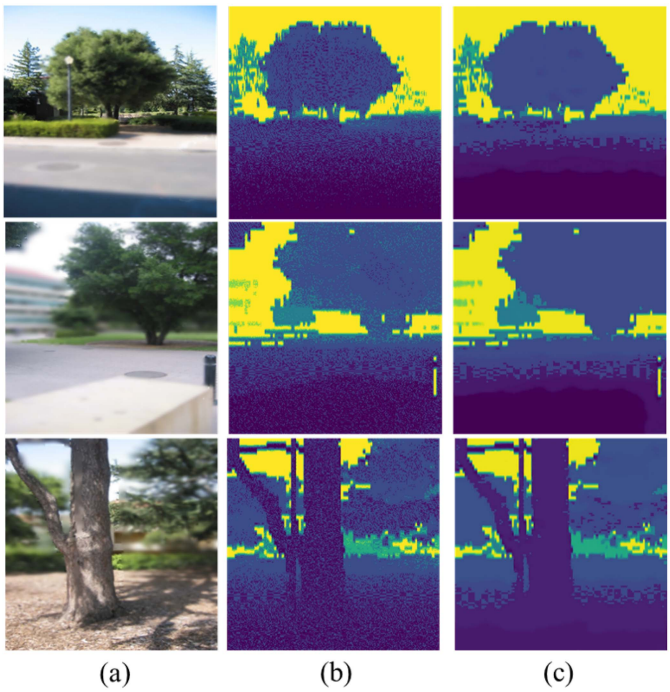

Fig. 7. 2HDED:NET results for depth estimation on Make3D dataset using the default loss L_{2HDED}: (a) RGB defocused image (b) Depth ground truth (c) Estimated depth.

contribute together to the training of the network as long as the loss function combines the depth and deblurring errors. Nevertheless, the network could be trained and still run by keeping a single head.

In this subsection, we evaluate the benefits of the two-headed architecture to the overall accuracy of the model. Hence, we alternately remove one head and retrain the network by using either the AiF or depth error depending on what head is preserved. Table III shows the results. For image deblurring, when the depth head is ablated and the loss $L_{charb} + \Psi(1 - SSIM)$ is used, the PSNR decreases by almost 3 dB, from 34.849 dB to 31.941 dB. Similarly, by ablating the deblurring head, the accuracy of the estimated depth maps becomes worse. On average, the RMSE increases by 0.05.

Thus, it is clear that 2HDED:NET achieves the best results when both heads are used together. Each head improves the results of the other one by complementing the ground truth, even if it is of a different nature.

C. Comparison With SoA Methods for Depth Estimation and Image Deblurring

We compare 2HDED:NET with some state-of-the-art solutions based on neural networks for depth estimation and image deblurring. Since in the literature, there are very few networks that solve simultaneously the problems of DFD and image deblurring [10], [29], we also consider recent methods dedicated exclusively to depth estimation. The blur is rarely taken into consideration in such cases [50], [51], [52], most of the networks being trained on AiF images. Table IV presents in the left half, results for depth estimation obtained with networks trained on NYU and Make3D dataset.

For NYU dataset, the best accuracy in terms of RMSE is obtained by Carvalho et al. [11] and Song et al. [13], both trained on defocused images with the sole purpose of generating depth maps. Their performances are very close, [13] outperforms [11] on $Abs.rel$ but not on RMSE. The approach proposed in [13] demonstrates improvement in performance by utilizing pairs of images with varying degrees of defocus to estimate depth, thereby providing additional ground truth information.

From the same category of networks using defocused i-mages, there are [29] and [10]. They are the most representative for our comparison since these networks handle both depth maps and blurred images. On average, the depth maps accuracy of [10] is worse by 0.2 in RMSE comparing with the best result in [11]. Gur et al. [29] lags behind with an RMSE of 0.766 but the results are still remarkable given the fact that they use self-supervised learning. 2HDED:NET is at half way between [11] and [10] with a RMSE of 0.241. In the category of networks handling both depth and deblurred images, our 2HDED:NET is the best in all metrics.

We also present results for three recent networks trained on AiF images to generate exclusively depth maps. The average RMSE ranges between 0.433 and 0.579, well inferior to the results of [11] or [13] and to our result. This difference proves the effectiveness of the defocus in the training set. The defocus is an additional source of information, independent of the scene geometry, which is commonly exploited by neural networks. Fig. 6 depicts three examples of depth maps obtained with 2HDED:NET. The visual comparison with the ground truth shows high-quality results. The smoothing regularization added to $L1$ loss instructs the network to produce depth maps with sharp edges and smooth, homogeneous regions that match well the ground truth and have significantly fewer artifacts.

Fig. 8. 2HDED:NET results for deblurring on NYU-v2 and Make3D dataset. From left to right: (a) defocused image, (b) ground truth AiF image, and (c) deblurred image. Similarly, (d), (e), and (f) for a different scene. Zoomed-in patches are shown below each scene.

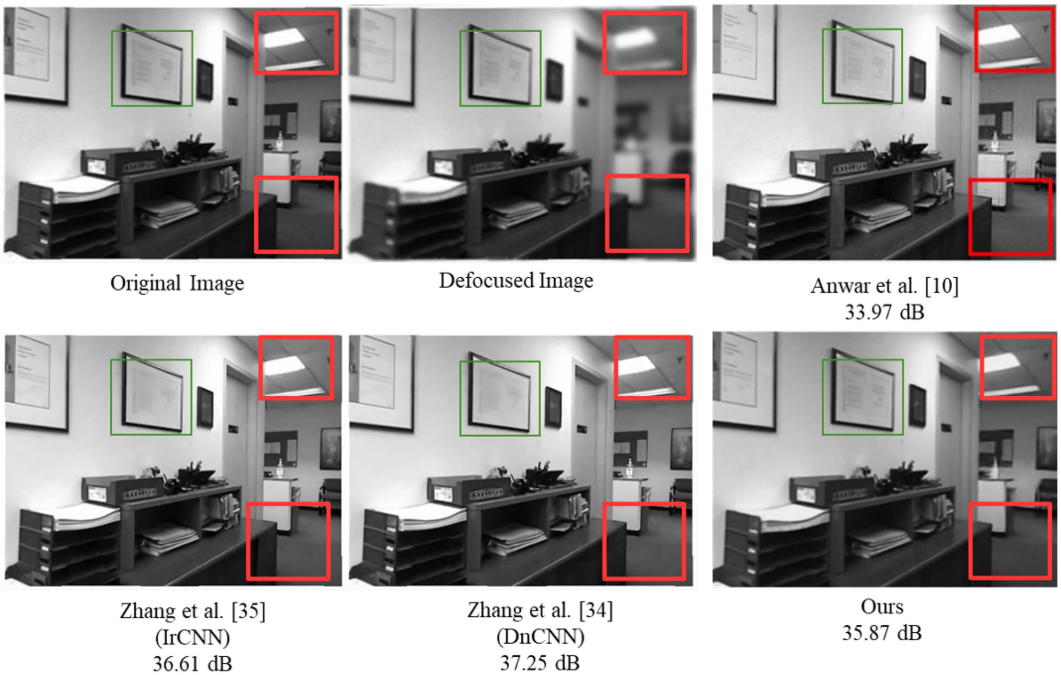

Fig. 9. Comparison of 2HDED:NET with the pipeline solution of Anwar et al. [10], and two general methods for image restoration [34], [35]: an example from NYU-v2 dataset.

TABLE IV
COMPARISON OF 2HDED:NET WITH SOA METHODS FOR DEPTH ESTIMATION AND IMAGE DEBLURRING ON NYU-V2 AND MAKE3D DATASETS

			Depth Estimation					Deblurring	
NYU dataset									
Method	Depth	Deblur	*RMSE* ↓	*Abs. rel* ↓	$\delta(1)$ ↑	$\delta(2)$ ↑	$\delta(3)$ ↑	*PSNR* ↑	*SSIM* ↑
Tang et al. [50]	✓	×	0.579	0.132	0.826	0.936	0.992	–	–
Chang et al. [51]	✓	×	0.433	0.087	**0.930**	**0.990**	**0.999**	–	–
Dong et al. [52]	✓	×	0.537	0.146	0.799	0.951	0.988	–	–
Song et al. [13]	✓	×	0.154	**0.028**	–	–	–	–	–
Carvalho et al. [11]	✓	×	**0.144**	0.036	–	–	–	–	–
Gur et al. [29]	✓	×	0.766	0.255	0.691	0.880	0.944	–	–
Zhang et al. [34] (DnCNN)	×	✓	–	–	–	–	–	32.43	0.67
Zhang et al. [35] (IrCNN)	×	✓	–	–	–	–	–	**35.46**	**0.99**
Anwar et al. [10]	✓	✓	0.347	0.094	–	–	–	34.21	–
2HDED:Net	✓	✓	0.244	0.029	0.914	0.979	0.995	34.85	0.99
Make3D dataset									
			C1-Error		C2-Error				
	Depth	Deblur	*RMSE* ↓	*Abs. rel* ↓	*RMSE* ↓	*Abs. rel* ↓		*PSNR* ↑	*SSIM* ↑
Fu et al. [12]	✓	×	**3.970**	0.157	7.32	**0.162**		–	–
Gur et al. [29]	✓	×	8.822	0.568	10.147	0.575		–	–
Zhang et al. [34] (DnCNN)	×	✓	–	–	–	–		23.16	0.68
Zhang et al. [35] (IrCNN)	×	✓	–	–	–	–		**27.09**	0.70
2HDED:Net	✓	✓	4.178	**0.153**	**6.132**	0.170		24.76	**0.78**

We also evaluate the performance of 2HDED:NET on Make3D dataset, which consists of outdoor scenes. The common approach to measuring the depth accuracy on this data set is to estimate errors on two depth ranges: C1 for depth up to 70 m and C2 up to 80m [53]. The results for Make3D dataset are shown in the lower half of Table IV. Earlier methods [12], [29], trained on Make3D dataset, are also displayed for comparison. Our analysis shows that 2HDED:NET performs better than [29] in all metrics and for both ranges. In what concerns [12], our results are better in terms of $Abs.rel.$ on C1 range and $RMSE$ on C2.

Fig. 7 depicts qualitative results for three different scenes. 2HDED:NET manages to correctly extract the depth of both near and distant regions.

Regarding image deblurring, in order to have a fair comparison, we selected only methods that are tested on NYU and/or Make3D benchmarks. The number of such methods is limited as these benchmarks are typically employed for depth estimation, rather than deblurring. Thus, for the NYU dataset, we selected [10], [29], [54] as a baseline, and for the Make3D dataset, only [10]. The results are shown in the right half of Table IV.

2HDED:NET performs better than the selected methods for NYU dataset, where it achieves an average PSNR of 34.85 dB. Comparing with [10], which is the main competitor, our results are superior by a margin of 0.64 dB. 2HDED:NET outperforms this method also on Make3D dataset, where the average PSNR is higher by 3 db.

Some qualitative results are depicted in Fig. 8. On the first row, there are two scenes from the NYU dataset. For each scene, the AiF image 8(b), its artificially defocused counterpart 8(a), and the deblurred version output by 2HDED:NET 8(c) can be compared. It can be seen how the blur is reduced both from near and far ranges. Two small areas, one with the stickers on the fridge in the foreground and the other with the margin of the hob closer to the camera are zoomed in on the second row. Fig. 8(c) shows how the details come to the surface after deblurring in both distant and near areas. 2HDED:NET works similarly on the second scene depicted in Fig. 8(d)–(f). Far and near-range patches are zoomed in order to prove the quality of the restored details.

For the outdoor scenes in the Make3D dataset, the quality of the restored image can be observed from the two examples in Fig. 8(g)–(l). The street light and the manhole cover that appear highly defocused in Fig. 8(g) are well restored in Fig. 8(i). Similarly, the tree branches in the second scene of Fig. 8(j) are obviously restored by 2HDED:NET in Fig. 8(l).

We also compared our results with the two networks, IrCNN and DnCNN, proposed by Zhang Kai et al. in [34] and [35] for deblurring. Since the reported results were for other benchmarks, we retrained and tested the networks on our datasets. The DnCNN is under 2HDED:NET in PSNR and SSIM of the deblurred images, while IrDNN overcomes our solution by 0.6 dB on average on the NYU dataset, and by 2.33 dB on Make3D. Still, there are cases like the image in Fig. 10 with fine textures, where our network performs much better (a PSNR gain of 2 dB). It seems that the fine textures are better restored by 2HDED:NET.

In the evaluation of 2HDED:NET, we placed particular emphasis on comparing it with the network proposed by Anwar et al. [10]. Although both networks provide depth maps and deblurred images, Anwar's network utilizes a pipeline processing approach, making it a suitable point of comparison with our network.

In Fig. 9, we give an example of an image from the NYU dataset restored by both [10] and 2HDED:NET. Our method achieves a PSNR of 35.87 dB, which is almost 2 dB higher than that of [10] 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. Another example, from the outdoor Make3D dataset, is depicted in Fig. 10. In this particular image, 2HDED:NET archives a PSNR of 37.19 dB,

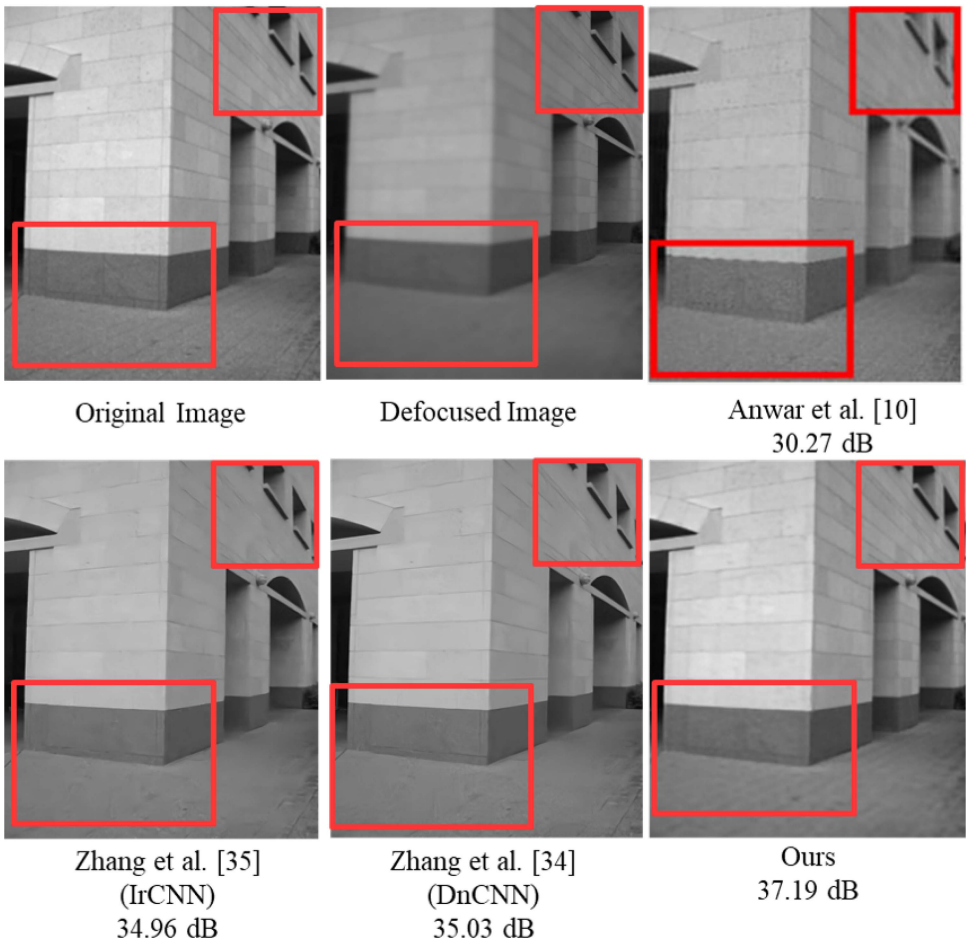

Fig. 10. Comparison of 2HDED:NET with the pipeline solution of Anwar et al. [10], and two general methods for image restoration [34], [35]: an example from Make3D dataset.

which is higher by almost 7 dB when compared to that obtained by [10]. The highly textured area of the wall, with tiles that are almost invisible in the image obtained by [10], is well restored in the image output by 2HDED:NET. Another advantage of 2HDED:NET with respect to [10] is the speed of computation. Once trained, 2HDED:NET generates the restored images very quickly, while in the case of [10], the restoration is a long process because of pixel-wise non-blind deconvolution.

VI. Conclusion and Future Work

In this work, we presented a novel deep convolutional neural network that estimates depth and restores the AiF images from a single out-of-focus image. The proposed network has a two-headed architecture consisting of an encoder and two parallel decoders, each of which with different roles: one outputs the depth map and the other the deblurred image. The formulation of an architecture that estimates the depth maps while removing blur from out-of-focus images, distinguishes our network from existing methods that are using pipeline processing. By parallelizing the tasks, the complexity of the network is reduced, while the depth estimation and blur removal work together toward performances that prove to be superior or close to the state-of-the-art results. Extensive tests on indoor and outdoor benchmarks have shown that 2HDED:NET outperforms the existing pipeline networks in both DFD and image deblurring. For the novel architecture of 2HDED:NET, we have proposed a new loss function that fuses depth and AiF errors, traditionally used separately in deep learning.

Since we experimented with synthetically blurred datasets, our future work will focus on developing a real defocused dataset containing depth ground truth, AiF and naturally defocused images.

References

[1] C. Godard, O. M. Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 270–279.

[2] A. Saxena, M. Sun, and A. Y. Ng, "Make3D: Learning 3D scene structure from a single still image," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 5, pp. 824–840, May 2009.

[3] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, vol. 27.

[4] C. Tang, C. Hou, and Z. Song, "Depth recovery and refinement from a single image using defocus cues," *J. Modern Opt.*, vol. 62, no. 6, pp. 441–448, 2015.

[5] A. Levin, R. Fergus, F. Durand, and W. T. Freeman, "Image and depth from a conventional camera with a coded aperture," *ACM Trans. Graph.*, vol. 26, no. 3, pp. 70-es, 2007.

[6] M. Masoudifar and H. Pourreza, "Depth estimation and deblurring from a single image using an optimized-throughput coded aperture," *J. Elect. Comput. Eng. Innovations*, vol. 11, 51–64, 2022.

[7] C. Hazirbas, S. G. Soyer, M. C. Staab, L. Leal-Taixé, and D. Cremers, "Deep depth from focus," in *Proc. Asian Conf. Comput. Vis.*, 2018, pp. 525–541.

[8] M.-B. Lien et al., "Ranging and light field imaging with transparent photodetectors," *Nature Photon.*, vol. 14, no. 3, pp. 143–148, 2020.

[9] Z. Huang, J. A. Fessler, T. B. Norris, and I. Y. Chun, "Light-field reconstruction and depth estimation from focal stack images using convolutional neural networks," in *Proc. IEEE ICASSP Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 8648–8652.

[10] S. Anwar, Z. Hayder, and F. Porikli, "Deblur and deep depth from single defocus image," *Mach. Vis. Appl.*, vol. 32, no. 1, pp. 1–13, 2021.

[11] M. Carvalho, B. L. Saux, P. Trouvé-Peloux, A. Almansa, and F. Champagnat, "Deep depth from defocus: How can defocus blur improve 3D estimation using dense neural networks?," in *Proc. Eur. Conf. Comput. Vis. Workshops*, 2018.

[12] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2002–2011.

[13] G. Song and K. M. Lee, "Depth estimation network for dual defocused images with different depth-of-field," in *Proc. IEEE 25th Int. Conf. Image Process.*, 2018, pp. 1563–1567.

[14] L. Ruan, B. Chen, J. Li, and M.-L. Lam, "AIFNet: All-in-focus image restoration network using a light field-based dataset," *IEEE Trans. Comput. Imag.*, vol. 7, pp. 675–688, 2021.

[15] A. Abuolaim and M. S. Brown, "Defocus deblurring using dual-pixel data," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 111–126.

[16] K. Zhang et al., "Deep image deblurring: A survey," *Int. J. Comput. Vis.*, vol. 130, no. 9, pp. 2103–2130, 2022.

[17] C. Li, "A survey on image deblurring," 2022, *arXiv:2202.07456*.

[18] R. Caruana, "Multitask learning," *Mach. Learn.*, vol. 28, no. 1, pp. 41–75, 1997.

[19] S. Ruder, "An overview of multi-task learning in deep neural networks," 2017, *arXiv:1706.05098*.

[20] M. Crawshaw, "Multi-task learning with deep neural networks: A survey," 2020, *arXiv:2009.09796*.

[21] Y. Lu, M. Sarkis, and G. Lu, "Multi-task learning for single image depth estimation and segmentation based on unsupervised network," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2020, pp. 10788–10794.

[22] Y. Wang, Y.-H. Tsai, W.-C. Hung, W. Ding, S. Liu, and M.-H. Yang, "Semi-supervised multi-task learning for semantics and depth," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2022, pp. 2505–2514.

[23] R. Xu, Z. Xiao, J. Huang, Y. Zhang, and Z. Xiong, "EDPN: Enhanced deep pyramid network for blurry image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 414–423.

[24] K. Xian, J. Zhang, O. Wang, L. Mai, Z. Lin, and Z. Cao, "Structure-guided ranking loss for single image depth prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 611–620.

[25] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, "Deeper depth prediction with fully convolutional residual networks," in *Proc. IEEE 4th Int. Conf. 3D Vis.*, 2016, pp. 239–248.

[26] Y. Cao, Z. Wu, and C. Shen, "Estimating depth from monocular images as classification using deep fully convolutional residual networks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 11, pp. 3174–3182, Nov. 2018.

[27] H. Jung, Y. Kim, D. Min, C. Oh, and K. Sohn, "Depth prediction from a single image with conditional adversarial networks," in *Proc. IEEE Int. Conf. Image Process.*, 2017, pp. 1717–1721.

[28] M. Carvalho, B. L. Saux, P. Trouvé-Peloux, A. Almansa, and F. Champagnat, "On regression losses for deep depth estimation," in *Proc. IEEE 25th Int. Conf. Image Process.*, 2018, pp. 2915–2919.

[29] S. Gur and L. Wolf, "Single image depth estimation trained via depth from defocus cues," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 7683–7692.

[30] S. Nah, T. H. Kim, and K. M. Lee, "Deep multi-scale convolutional neural network for dynamic scene deblurring," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3883–3891.

[31] X. Tao, H. Gao, X. Shen, J. Wang, and J. Jia, "Scale-recurrent network for deep image deblurring," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8174–8182.

[32] O. Kupyn, V. Budzan, M. Mykhailych, D. Mishkin, and J. Matas, "Deblurgan: Blind motion deblurring using conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8183–8192.

[33] J. Lee, H. Son, J. Rim, S. Cho, and S. Lee, "Iterative filter adaptive network for single image defocus deblurring," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 2034–2042.

[34] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3142–3155, Jul. 2017.

[35] K. Zhang, W. Zuo, S. Gu, and L. Zhang, "Learning deep CNN denoiser prior for image restoration," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3929–3938.

[36] Z. Zha, B. Wen, X. Yuan, J. T. Zhou, J. Zhou, and C. Zhu, "Triply complementary priors for image restoration," *IEEE Trans. Image Process.*, vol. 30, pp. 5819–5834, 2021.

[37] Z. Zha, X. Yuan, J. Zhou, C. Zhu, and B. Wen, "Image restoration via simultaneous nonlocal self-similarity priors," *IEEE Trans. Image Process.*, vol. 29, pp. 8561–8576, 2020.

[38] G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4700–4708.

[39] X. Zhang, F. Wang, H. Dong, and Y. Guo, "A deep encoder-decoder networks for joint deblurring and super-resolution," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 1448–1452.

[40] S. Nazir, L. Vaquero, M. Mucientes, V. M. Brea, and D. Coltuc, "2HDED: Net for joint depth estimation and image deblurring from a single out-of-focus image," in *Proc. IEEE Int. Conf. Image Process.*, 2022, pp. 2006–2010.

[41] J. Hu, M. Ozay, Y. Zhang, and T. Okatani, "Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2019, pp. 1043–1051.

[42] C. Godard, O. M. Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 3828–3838.

[43] R. Timofte, R. Rothe, and L. V. Gool, "Seven ways to improve example-based single image super resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1865–1873.

[44] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2017, pp. 136–144.

[45] P. Charbonnier, L. Blanc-Feraud, G. Aubert, and M. Barlaud, "Two deterministic half-quadratic regularization algorithms for computed imaging," in *Proc. IEEE 1st Int. Conf. Image Process.*, 1994, vol. 2, pp. 168–172.

[46] A. Abuolaim, R. Timofte, and M. S. Brown, "Ntire 2021 challenge for defocus deblurring using dual-pixel images: Methods and results," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 578–587.

[47] S. Nazir and D. Coltuc, "Edge-preserving smoothing regularization for monocular depth estimation," in *Proc. IEEE 26th Int. Conf. Automat. Comput.*, 2021, pp. 1–6.

[48] J. Lee, S. Lee, S. Cho, and S. Lee, "Deep defocus map estimation using domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 12222–12230.

[49] M. Potmesil and I. Chakravarty, "A lens and aperture camera model for synthetic image generation," *ACM SIGGRAPH Comput. Graph.*, vol. 15, no. 3, pp. 297–305, 1981.

[50] M. Tang, S. Chen, R. Dong, and J. Kan, "Encoder-decoder structure with the feature pyramid for depth estimation from a single image," *IEEE Access*, vol. 9, pp. 22640–22650, 2021.

[51] J. Chang and G. Wetzstein, "Deep optics for monocular depth estimation and 3D object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 10193–10202.

[52] X. Dong, M. A. Garratt, S. G. Anavatti, and H. A. Abbass, "Mobilexnet: An efficient convolutional neural network for monocular depth estimation," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 11, pp. 20134–20147, 2022.

[53] M. Liu, M. Salzmann, and X. He, "Discrete-continuous depth estimation from a single image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 716–723.

[54] J. Qiu, X. Wang, S. J. Maybank, and D. Tao, "World from blur," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 8493–8504.

Saqib Nazir (Graduate Student Member, IEEE) received the bachelors's and master's degrees in computer science from COMSATS University Islamabad, Pakistan. He is currently working toward the Ph.D. degree with the CEOSpace Tech. of Polytechnic University of Bucharest, Romania. His role as an Early Stage Researcher with the MENELOAS-NT Project is to acquire depth from defocus images using state-of-the-art deep learning methods. His research interests include computer vision, image analysis, and machine & deep learning.

Lorenzo Vaquero received the B.S. degree in computer science in 2018, and the M.S. degree in Big Data in 2019. He is currently working toward the Ph.D. degree with the CiTIUS, University of Santiago de Compostela, Santiago, Spain. His research interests include visual object tracking and deep learning for autonomous vehicles.

Víctor M. Brea is currently an Associate Professor with CiTIUS, University of Santiago de Compostela, Santiago, Spain. He has authored more than 100 scientific papers in these fields of research. His main research interests include computer vision, both on deep learning algorithms, and on the design of efficient architectures and CMOS solutions.

Manuel Mucientes is currently an Associate Professor with the CiTIUS of the University of Santiago de Compostela, Spain. He has authored more than 100 scientific papers in his research field, which include artificial intelligence applied to the following areas: computer vision for object detection and tracking, machine learning, and process mining.

Daniela Coltuc received the M.Sc. and Ph.D. degrees and the Habilitation degree in electronics, telecommunications and information technology from Univ. POLITEHNICA of Bucharest, Bucharest, Romania. She is currently a Full Professor with this university. She was also an Invited Professor with Jean Monnet University, St. Etienne and University de Lyon, France. She has a sound background in information theory with applications in image processing. In the recent years, she has worked in computational imaging.