2HDED:NET FOR JOINT DEPTH ESTIMATION AND IMAGE DEBLURRING FROM A SINGLE OUT-OF-FOCUS IMAGE

Saqib Nazir[1,2]*, Lorenzo Vaquero*[2]*, Manuel Mucientes*[2]*, Víctor M. Brea*[2]*, Daniela Coltuc*[1]

[1]CEOSpaceTech, University POLITEHNICA of Bucharest (UPB), Bucharest, Romania
[2]Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS)
University of Santiago de Compostela (USC), Santiago de Compostela, Spain

ABSTRACT

Depth estimation and all-in-focus image restoration from defocused RGB images are related problems, although most of the existing methods address them separately. The few approaches that solve both problems use a pipeline processing to derive a depth or defocus map as an intermediary product that serves as a support for image deblurring, which remains the primary goal. In this paper, we propose a new Deep Neural Network (DNN) architecture that performs in parallel the tasks of depth estimation and image deblurring, by attaching them the same importance. Our Two-headed Depth Estimation and Deblurring Network (2HDED:NET) is an encoder-decoder network for Depth from Defocus (DFD) that is extended with a deblurring branch, sharing the same encoder. The network is tested on NYU-Depth V2 dataset and compared with several state-of-the-art methods for depth estimation and image deblurring.

Index Terms— Depth from Defocus, Image Deblurring, Deep learning

1. INTRODUCTION

Over the last few years, there has been a lot of interest in depth estimation and all-in-focus image restoration from a single image. The depth estimation is critical for scene understanding and for 3D reconstruction in robotics, augmented reality or autonomous driving and flight [1]. No less important, defocus deblurring is an essential part of applications like face and object recognition or image segmentation and classification, to name a few [2].

Defocus blur occurs in images taken with a shallow Depth of Field (DoF) caused by large aperture sizes. Although extensively studied in the past and rigorously modelled, the defocus blur remains difficult to estimate in applications because it varies not only with the distance to the object but also spatially. Image deblurring techniques aim to reduce the blur and to restore a sharp image from its defocused version.

In depth estimation, the defocus blur is one of the cues used by the existing methods. Over the years, numerous efforts have been made to estimate depth from single defocused images [1, 3] but despite its valuable content, the defocus blur cannot be used solely for depth estimation. There are needed supplementary constraints like coded apertures[4], dual images[5], multi focus stack[6] etc. The advent of deep Convolution Neural Networks (CNN) has taken the performance of Single Image Depth Estimation (SIDE) and image deblurring to the next level. Most of the CNN-based solutions consider these two problems separately, being dedicated to either depth estimation or image deblurring. The few networks that treat them jointly use a pipeline processing to derive first a depth or defocus map that serves next as support for image deblurring [2, 7]. It is known from the literature that concatenating two networks generally increases the complexity and may deteriorate the model for certain tasks [8]. To address this issue, we formulate a two-fold task that combines depth estimation and image deblurring by attaching them the same importance. To fulfill the task in a balanced way, we propose a new deep CNN, the two-Headed Depth Estimation, and defocus Deblurring NETwork (2HDED:NET). The network consists of an encoder and a decoder splitted in two branches that work in parallel for Depth from Defocus (DFD) and image deblurring. The branches interact with each other during the training, enabling the encoder to learn semantically rich features that are well suited for both tasks. Unlike the pipelined solutions, the architecture of 2HDED:NET is straightforward, simple, and easy to train. A distinctive feature of 2HDED:NET is that once fully trained, the depth estimation branch is no longer necessary to recover all-in-focus image and vice versa.

The main contributions of our work are:

- A novel architecture 2HDED:NET that recovers the all-in-focus images and generates the depth map from a single defocused image.
- The architecture is the first of its kind to generate the depth map and all in-focus images in a balanced way

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 860370. The last author acknowledges financial support from UEFISCDI Romania grant 31/01.01.2021 PN III, 3.6 Suport.

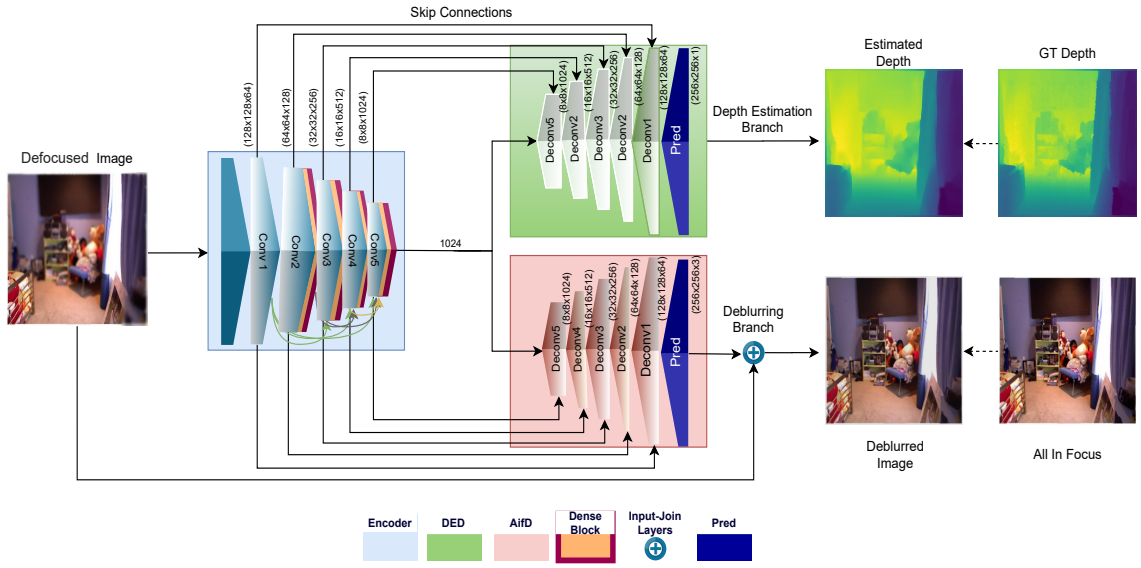

Fig. 1. 2HDED:NET architecture consists of one encoder and two decoders that work in parallel. The upper branch estimates the depth map and the lower one the all-in-focus image. The network is fed in with defocused RGB images.

and attach the same importance to both tasks.

- A new loss function that combines constraints from both depth estimation and deblurring and enforces the encoder to learn much richer semantic features.
- Experimental results on NYU-depth V2 dataset enriched with synthetic defocused images, which confirm the effectiveness of our approach.

2. RELATED WORK

This section briefly reviews several CNN based solutions for DFD and deblurring.

In depth estimation, the main feature that drives the learning is the scene geometry. The depth maps can be estimated from single images by using solely this kind of feature either in a supervised or self-supervised frame [9]. Cues like motion in videos or different perspectives in stereo vision leverage the geometry but in such cases, the network has to work with a sequence of frames or stereo pairs. Defocus is another cue that can be used to improve depth estimation. The amount of defocus in an image depends on how far an object is from the camera. Moreover, the defocus is generated by the camera, exists in any image, and consequently, is appropriate for SIDE. So far, the tasks of depth estimation from defocused images have been solved from dual images [5], focal stack [6], multi-view or stereo pairs [10]. Undoubtedly, to tackle the aforementioned task with a single image adds up another level of difficulty in achieving high accuracy. In this line of research, some deep architectures that significantly improve the performance of depth estimation have been recently proposed. Carvalho et al. [1] designed a supervised CNN to estimate depth in the wild using defocus blur. Anwar et al. [2] cascaded a CNN and a fully connected NN to estimate depth and next, an all-in-focus images from a single defocused image. The authors in [11] jointly explore Conditional Random Field (CRF) and deep CNN for DFD from a single image. The category of self-supervised learning is illustrated by the method outlined in [12], which introduces a framework for depth estimation using the DFD and depth-from-focus algorithms on defocus stacks, or by [3] that uses a Point Spread Function (PSF) convolutional layer to improve depth estimation using the defocus cue.

When the PSF of the camera is not known as it is the case in many applications, the deblurring is an ill-posed problem since it is required to retrieve from the defocused image both the blur kernel and the latent sharp image. In NN-based approaches, the common strategy is to first estimate a defocus map that subsequently guides the image deblurring. An example of such pipeline processing is [7], where an encoder-decoder is trained to estimate the defocus map that is fed together with the defocused RGB image into a fully convolutional encoder-decoder with skip connections that estimates a sharp image. The scheme is completed by a domain adaptation module, which is the discriminator of a Generative Adversarial Network (GAN). It minimizes the domain difference between the feature distributions in the training set, which is synthetic, and the natural defocused Light Field (LF) dataset. The method achieves good results in removing blur on the LF dataset but is complex as GANs are hard to train. Anwar et al. [2] use a similar approach and train an encoder-decoder network to estimate this time a depth map that is used next to compute kernels for reconstructing the all in-focus image.

3. 2HDED:NET

Figure 1 depicts the architecture of 2HDED:NET. The network consists of one encoder and two decoders designed to output depth and all in-focus images in parallel. By sharing the information learned by the same encoder, both branches can benefit from each other. In terms of complexity, 2HDED:NET is simpler than [7] that concatenates two encoder-decoder networks. 2HDED:NET is a supervised method, which means it needs for training ground truth depth as well as all in-focus images.

We build on DenseNet-121 [13] as the encoder. The reason is that this network reuses features by concatenating the features of each layer with those from the next layers instead of summing them. The goal of concatenation is to use the features obtained in the previous layers in the deeper layers as well. This is known as "feature reusability". DenseNets can learn mappings with fewer parameters than a typical CNN because there is no need to learn redundant maps. Similar to [1], we replace the max-pooling layer with a 4×4 convolution layer to reduce resolution while increasing the number of feature channel maps. We use skip connections between encoder and decoder parts to make the learning easier. The skip connections prevent the gradient vanishing problem while making the learning process easier, as subsequent layers focus on solving residuals rather than entirely new representations. The encoder helps in obtaining 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 [1]. It consists of 5 decoding blocks, each with a 4×4 transposed convolution that increases the resolution of the feature map, which is then followed by a 3×3 convolution that helps reducing the aliasing effect of the upsampling. Batch normalization and ReLU functions are included after each convolutional layer to make learning more stable and allow the representation of non-linearities.

As loss function for depth estimation, we resort to $L1$ norm that calculates the average of absolute difference between the estimated depth $\widehat{I}^{depth}$ and the ground truth I^{depth}:

$$L_{depth} = 1/n \sum_{t=1}^{n} |\widehat{I}_t^{depth} - I_t^{depth}| \quad (1)$$

This norm is appropriate for estimating sparse solutions as is the case of depth maps [14].

We refer to the deblurring decoder as the All-in-focus Decoder (AifD). Unlike the DED, the output of AifD is a three channel RGB image. We use an input-join layer to aggregate the blurred input image with the output of AifD like in [7, 8] for final prediction. The content of the defocused image and the corresponding prediction of AifD are embedded in the input-join layer, providing this branch with more detailed guidance to learn the deblurring. Unlike the methods that use a pipeline processing [2, 7], where the depth or defocus map is predicted first and the all-in-focus image is recovered subsequently, our deblurring branch is not built on such estimates, avoiding the dependence on unsatisfactory depth maps in some cases.

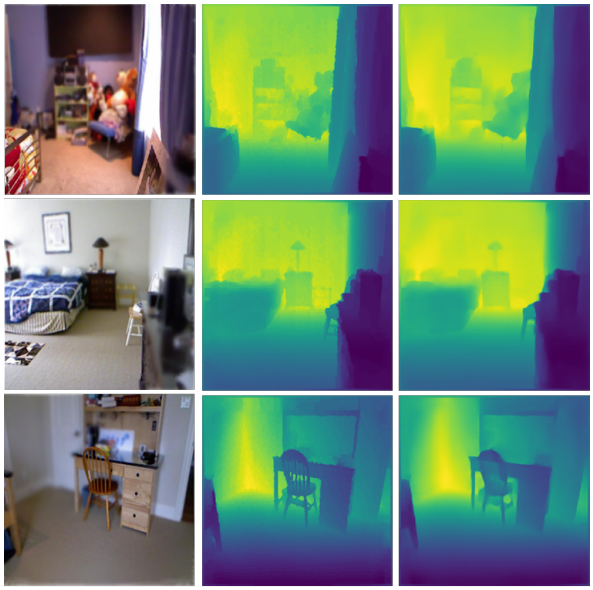

Fig. 2. 2HDED:NET results for depth estimation: (a) RGB image (b) Depth ground truth (c) Estimated Depth.

We include the Charbonnier loss function [15] to achieve robust regression and high-quality deblurring. This loss is calculated as a squared error between the estimated all-in-focus image $\widehat{I}^{aif}$ and the sharp image I^{aif}:

$$L_{deblur} = \frac{1}{W \cdot H} \sum_{i=1}^{W} \sum_{j=1}^{H} \sqrt{(\widehat{I}_{i,j}^{aif} - I_{i,j}^{aif})^2 + \epsilon^2} \quad (2)$$

where $W \times H$ is the image size and ϵ is a hyper-parameter set to $1e-3$. This hyper-parameter acts as a pseudo-Huber loss and smooths the error if it is smaller than ϵ. The overall loss function of 2HDED:Net is:

$$L = L_{depth} + \mu L_{deblur} \quad (3)$$

Table 1. Quantitative comparison of 2HDED:NET with state of the art methods.

DEPTH ESTIMATION				DEBLUR
Method	*RMS* ↓	*REL* ↓	*LOG 10* ↓	*PSNR* ↑
Song[5]	0.154	**0.028**	**0.012**	–
Carvalho[1]	**0.144**	0.036	0.016	–
Gur[3]	0.766	0.25	0.092	–
Anwar[2]	0.347	0.094	0.039	**34.21**
2HDED:Net	0.285	0.035	0.024	32.11

(a) (b) (c) (d) (e) (f)

Fig. 3. 2HDED:NET results for deblurring. From left to right: (a) defocused image, (b) ground truth all-in-focus image and (c) deblurred image. Similarly, (d), (e) and (f) for a different scene. Zoomed-in patches are shown below each.

with $\mu = 0.1$.

A noticeable feature of 2HDED:NET is that once the model is fully trained, we are still able to accomplish one task if the other branch is removed, e.g. we can perform DFD without AifD branch and vice versa.

4. EXPERIMENTAL RESULTS

We have run experiments on NYU-Depth V2 dataset [16], which contains approximately 230,000 pairs of aligned all-in-focus images and corresponding depth maps from a Kinect camera. We use the same split as [1, 2], i.e., 249 scenes are employed for training and 215 for testing. The images from Kinect camera serve as ground truth for depth branch and the all-in-focus images are ground truth for the deblurring branch. To feed in the network, we generate synthetic out-of-focus images by using the thin lens model as in [17].

We compare 2HDED:NET with several state-of-the-art methods in Table 1. For depth we selected [2], which uses a pipeline architecture to output depth and all-in-focus images. We considered also [1, 3, 5] that are exclusively dedicated to depth estimation. All of them are trained and tested on NYU-Depth V2 dataset. From the four selected methods only [2] outputs all-in-focus images. We use it as reference also for the deblurring. To quantitatively evaluate the depth maps accuracy, we calculate the Root Mean Square error (RMS), Relative Error (Rel) and Log10. For deblurring performance, we use Peak Signal to Noise Ratio (PSNR).

Table 1 collects the results for depth estimation in the left-most columns. Our 2HDED:NET outperforms for all evaluation metrics the pipeline solution [2]. Especialy in terms of RMSE, 2HDED:NET leads by a margin of 0.062. Our results are below those of [1, 5], but these are methods dedicated to depth estimation, without recovering all-in-focus images. The performance of [3] is the poorest but this is a self-supervised approach while all the others are supervised. Figure 2 shows three depth maps estimated by 2HDED:NET. The RGB image and the ground truth are displayed for comparison. It can be seen how the depth is well captured for both close and far away objects in the scene.

Table 1 also includes the deblurring performances of 2HDED:NET. The obtained all-in-focus images have an average PSNR over 32dB, which is a good score given the fact that they are recovered from defocused images with about 27dB. This result is inferior by 2dB to that of [2]. Figure 3 depicts results for the deblurring branch. On the first row, there are the whole scenes and beneath, crops with details. The blurred images and the ground truth are also displayed for comparison. 2HDED:NET clearly restores edges and textures. Details of the toys near the teddy bear missing in Figure 3 (a), reoccur in (c). Likewise, 2HDED:Net recovers in (f) the door frame that is missing in (d).

In terms of network complexity, 2HDED:Net is significantly lighter than the pipeline architecture of Anwar et al. [2]. The total number of parameters of our network is 41M, which is three times less than that of [2], with 138M.

5. CONCLUSION

In this paper, we propose a novel CNN architecture with two parallel decoders that estimate depth and recover all-in-focus images from a single out-of-focus image. The two headed architecture distinguishes our network from existing met–hods that use pipeline processing. The tasks parallelization reduces the network complexity, all while maintaining performances in depth estimation and deblurring comparable with state-of-art approaches. Experimental results on the NYU-depth V2 dataset show that 2HDED:NET outperforms the pipeline approach in estimating the depth. Our future work will focus on further ablation studies and improvement of loss function to obtain better deblurring and more accurate depth estimation.

6. REFERENCES

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

[2] Saeed Anwar, Zeeshan Hayder, and Fatih Porikli, "Deblur and deep depth from single defocus image," *Machine vision and applications*, vol. 32, no. 1, pp. 1–13, 2021.

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

[4] Harel Haim, Shay Elmalem, Raja Giryes, Alex M Bronstein, and Emanuel Marom, "Depth estimation from a single image using deep learned phase coded mask," *IEEE Transactions on Computational Imaging*, vol. 4, no. 3, pp. 298–310, 2018.

[5] Gwangmo Song and Kyoung Mu Lee, "Depth estimation network for dual defocused images with different depth-of-field," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 1563–1567.

[6] Maxim Maximov, Kevin Galim, and Laura Leal-Taixé, "Focus on defocus: bridging the synthetic to real domain gap for depth estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 1071–1080.

[7] Lingyan Ruan, Bin Chen, Jizhou Li, and Miu-Ling Lam, "Aifnet: All-in-focus image restoration network using a light field-based dataset," *IEEE Transactions on Computational Imaging*, vol. 7, pp. 675–688, 2021.

[8] Xinyi Zhang, Fei Wang, Hang Dong, and Yu Guo, "A deep encoder-decoder networks for joint deblurring and super-resolution," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 1448–1452.

[9] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow, "Digging into self-supervised monocular depth estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3828–3838.

[10] Ravi Garg, Vijay Kumar Bg, Gustavo Carneiro, and Ian Reid, "Unsupervised cnn for single view depth estimation: Geometry to the rescue," in *European conference on computer vision*. Springer, 2016, pp. 740–756.

[11] Fayao Liu, Chunhua Shen, and Guosheng Lin, "Deep convolutional neural fields for depth estimation from a single image," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 5162–5170.

[12] Yawen Lu, Garrett Milliron, John Slagter, and Guoyu Lu, "Self-supervised single-image depth estimation from focus and defocus clues," *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 6281–6288, 2021.

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

[14] Saqib Nazir and Daniela Coltuc, "Edge-preserving smoothing regularization for monocular depth estimation," in *2021 26th International Conference on Automation and Computing (ICAC)*. IEEE, 2021, pp. 1–6.

[15] Pierre Charbonnier, Laure Blanc-Feraud, Gilles Aubert, and Michel Barlaud, "Two deterministic half-quadratic regularization algorithms for computed imaging," in *Proceedings of 1st International Conference on Image Processing*. IEEE, 1994, vol. 2, pp. 168–172.

[16] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus, "Indoor segmentation and support inference from rgbd images," in *European conference on computer vision*. Springer, 2012, pp. 746–760.

[17] Junyong Lee, Sungkil Lee, Sunghyun Cho, and Seungyong Lee, "Deep defocus map estimation using domain adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12222–12230.