ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10448112

EFFICIENT FUSION OF DEPTH INFORMATION FOR DEFOCUS DEBLURRING

Jucai Zhai *Yang Liu* *Pengcheng Zeng* *Chihao Ma* *Xinan Wang*⋆ *Yong Zhao*⋆

School of Electronic and Computer Engineering, Shenzhen Graduate School
Peking University
Lishui Road 2199, Nanshan District, Shenzhen, China 518055

ABSTRACT

Defocus deblurring is a classic problem in image restoration tasks. The formation of its defocus blur is related to depth. Recently, the use of dual-pixel sensor designed according to depth-disparity characteristics has brought great improvements to the defocus deblurring task. However, the difficulty of real-time acquisition of dual-pixel images brings difficulties to algorithm deployment. This inspires us to remove defocus blur by single image with depth information. We propose a single-image depth-enhanced defocus deblurring network, which uses a depth map estimated by the monocular depth estimation network to guide the network defocus deblurring. We design a deep information fusion unit, which greatly improves the effect of deblurring. Experiments show that on the single image defocus deblurring task, the experimental results demonstrate the superiority of our method.

Index Terms— defocus deblurring, depth, deep information fusion

1. INTRODUCTION

Scene points that are outside the depth of field of the lens are out of focus when shooting with the camera [1]. This phenomenon is called defocus blur. The creation and removal of defocus blur has huge application scenarios, especially in the field of photography. Think about the magic of bokeh and sharpening people. The current method for defocus blurring is starting from a single blurred image to remove defocus blur [2, 3]. But it is still a challenge to remove large-scale blur.

Recent work [1] proposes a method to remove defocus blur using the left and right views of a dual-pixel (DP) sensor as input. This approach comes from the way sensors work, similar to stereo views [4] that provide disparity cues. DP sensor can capture defocus disparity with left and right views. Using this reliable disparity information, the amount of spatial blur can be estimated, reducing defocus blur. However, DP sensors are not widely deployed, and the inability to obtain real-time data limits its application. Some recent works have turned to use single-image to defocus deblurring. [5] proposed IFAN, [3] proposed KPAC, [6] and [7] utilized DP views to assist single-image defocus deblurring. We found that the above methods are still not ideal for the recovery of image details.

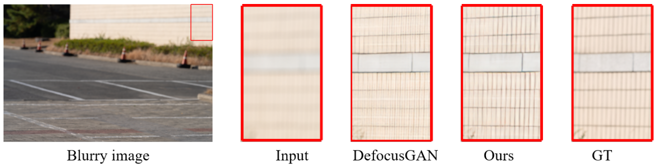

Fig. 1. Qualitative comparison of the generalization experiments on the RealDOF test set.

	PSNR↑	SSIM↑	Params(M)↓	MACs(G)↓
DefocusGAN [7]	24.08	0.723	4.59	281.2
Ours (1-branch)	24.35	0.729	2.65	163.9

Table 1. Quantitative comparison of the generalization experiments on the RealDOF test set.

That being the case, is it more efficient to use depth information directly to remove defocus blur? We know that the disparity information provided by the DP sensor is related to depth. Depth can assist in the elimination of defocus blur. At present, the combination of massive data and deep learning provides powerful prior information for monocular depth estimation [8]. Depth estimation is a channel for understanding 3D information from 2D images [9]. Methods for monocular depth estimation with self-supervised learning, such as Monodepth [10] and MonodepthV2 [11], have good scalability and generalization. So we use the depth map estimated by monocular assisted defocus deblurring is feasible in the idea.

Based on these findings, we propose a single-image depth-enhanced defocus deblurring network to alleviate the defocus blur problem. Especially, we propose an efficient fusion module to fuse depth and blur information for effective single-image defocusing deblurring, which is mainly composed of gated recurrent unit (GRU) [12]. Partitioning

⋆ Corresponding authors

This work is supported by the Shenzhen Science and Technology Research and Development Fund (No. JSGG202011022153800002, KQTD20200820113105004) and Shenzhen Science and Technology Program (No. JCYJ20180503182133411, JCYJ20220818102215034).

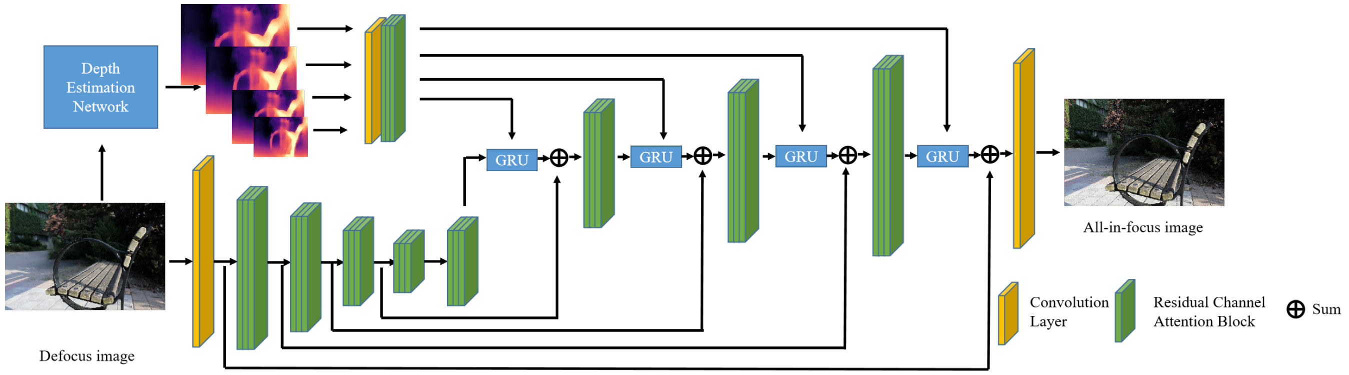

Fig. 2. Illustration of the proposed depth-enhanced single-image defocusing deblurring network. The network consists of three parts, blur feature extraction module, deep feature encoding module, and GRU fusion decoding module. The proposed network can make good use of depth information for defocus deblurring.

deblurring according to depth information can guide the network to deal with the amount of blurring in different regions. In Table 1, we compare the generalization performance of the proposed network with state-of-the-art network DefocusGAN [7] on the RealDOF test set [5]. The proposed method has good generalization performance. As shown in Figure 1, the proposed method performs well on walls with varying depths and can restore the texture of tiles.

Our main contributions are summarized as:

- We propose a single-image depth-enhanced defocus deblurring network that effectively exploits depth information for defocus deblurring compared to previous methods.
- Experimental results show that the proposed method is effective with a small number of parameters and performs well in the single-image defocus deblurring task.

2. PROPOSED METHOD

We propose a single-image depth-enhanced defocus deblurring network. The network is divided into three parts, (a) blur feature encoding module, (b) deep feature encoding module, (c) GRU fusion decoding module. Figure 2 shows the illustration of the depth-enhanced defocus deblurring network.

2.1. Overall Pipeline

Given an image $I_B \in R^{H\times W\times 3}$, first, use the convolutional layer to extract the shallow feature $F_0 \in R^{H\times W\times C_1}$. Then through 4 groups of residual channel attention blocks (RCAB), gradually downsample to $\frac{1}{16}$ of the original resolution. Complete the feature extraction and encoding work, during which the number of channels remains the same. Then use the monocular depth estimation network (MonodepthV2 [11] is used here) to estimate the depth map. Downsample the depth map to $\frac{1}{16}$ the original resolution step by step, and get 4 sets of depth feature vectors with channel C_2 through the same depth encoder. Finally, use GRU and RCAB to fuse and upsample the depth and blur feature vectors step by step, use the residual to connect the previous blur feature vectors of the same resolution, and finally use the decoder to restore the feature vectors to defocus deblurred images $I_{DB} \in R^{H\times W\times 3}$. In the experiment, we set C_1 and C_2 to 64 and 16 respectively.

2.2. Blur feature encoding module

We input the blurred image $I_B \in R^{H\times W\times 3}$ into the network, encode the image using a convolutional layer, and obtain shallow features $F_0 \in R^{H\times W\times 64}$. To obtain richer representations, we use residual channel attention blocks (RCAB) instead of convolutional layers to process shallow features. RCAB first uses average pooling to obtain the channel information of the feature and then multiplies the feature after convolution to realize the channel attention operation. Residual addition is then performed to implement the residual channel attention operation. Here we use 3 RCABs cascades to extract rich feature information. Then downsample it to $\frac{1}{2}$ the original resolution. Repeat this operation to obtain feature blocks with $\frac{1}{4}$, $\frac{1}{8}$, and $\frac{1}{16}$ resolutions in sequence.

2.3. Deep feature encoding module

Our innovation is to use depth information to guide defocus deblurring. First, we need to get the depth map. Here we use MonodepthV2 for depth map generation. Since MonodepthV2 is trained in an unsupervised method, it is better in generalization.

After obtaining the depth map, we upsample the depth map to the full resolution of the blurred image. Then downsample to $\frac{1}{2}$, $\frac{1}{4}$, $\frac{1}{8}$ of the resolution in turn, and align with the resolution of the features extracted from the blurred image. We use the deep feature extraction module to extract the fea-

Method	Indoor			Outdoor			Indoor & Outdoor				
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	MAE↓	LPIPS↓	Params(M)
JNB [13]	25.52	0.784	0.188	21.16	0.632	0.274	23.28	0.706	0.049	0.232	-
EBDB [2]	25.83	0.790	0.326	21.21	0.631	0.407	23.47	0.708	0.049	0.368	-
DMENet [14]	25.70	0.789	0.315	21.51	0.655	0.402	23.55	0.720	0.049	0.360	26.94
DPDNet(single) [1]	26.52	0.828	0.179	22.08	0.689	0.229	24.25	0.757	0.044	0.204	35.25
IFAN [5]	27.80	0.856	0.131	22.70	0.719	0.179	25.18	0.786	0.041	0.156	10.48
KPAC [3]	28.02	0.852	0.129	22.64	0.702	0.190	25.26	0.774	0.041	0.161	**2.06**
MDPNet [6]	28.02	0.840	0.186	22.82	0.689	0.261	25.35	0.763	0.040	0.225	46.86
DefocusGAN [7]	**28.31**	0.857	0.086	22.94	0.718	0.135	25.56	0.786	0.039	0.111	4.59
Ours (2-branch)	28.29	**0.861**	**0.084**	**23.07**	**0.721**	**0.134**	**25.61**	**0.789**	**0.038**	**0.109**	4.11

Table 2. Quantitative comparisons with single-image defocus deblurring methods. The best results are indicated in boldface. Results are on the DPDD dataset. (test set consists of 37 indoor and 39 outdoor scenes.)

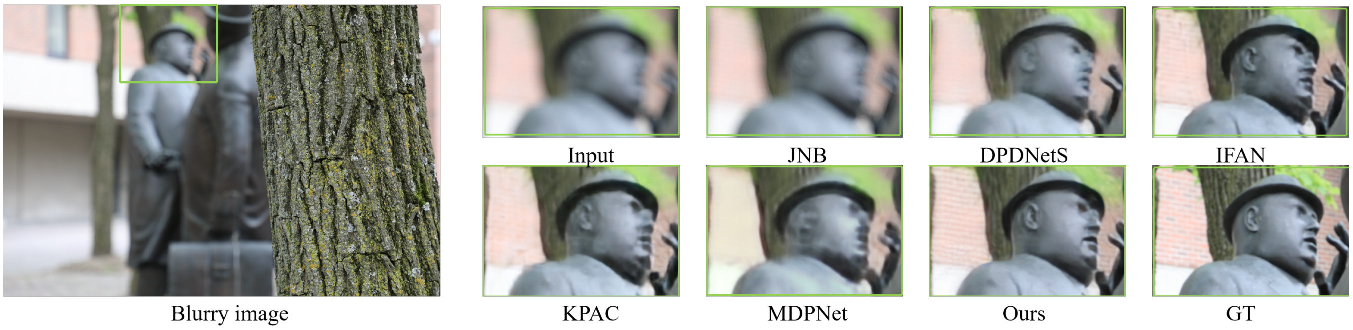

Fig. 3. Qualitative comparison with single-image defocus deblurring methods on the DPDD dataset. We show the deblurring results of different methods.

ture information of the depth map. The depth extraction module includes a convolutional layer and 3 RCABs for extracting depth information. The input here is $I_{Depth} \in R^{h \times w \times 3}$, output as $F_{Depth} \in R^{h \times w \times 16}$.

2.4. GRU fusion decoding module

After obtaining blur features and depth features, since defocus blur is spatially variable and related to depth, we need to consider how to effectively fuse depth information to achieve good defocus deblurring. Due to the unique gating mechanism, GRU [12] has better feature fusion characteristics with few parameters. Therefore, through the update and forgetting mechanism of GRU, we can achieve the matching of depth information and blur information, and flexibly deal with blur information in different regions. The use of GRU in stereo matching demonstrates its superiority [15]. We feed the depth information as the hidden state h. Let the blur feature be the input x of the GRU node, Then get the gating signal z, Then we splice h and x, update the hidden state through the $tanh$ activation function, and get new state h'. Finally, we use the gating signal update to get the output y.

To better deal with large-scale blurring, we use GRU deep fusion module to perform deblurring decoding operations on blurred features at different resolutions. We first perform residual cascade operation on the low-resolution blur features and then input 3 RCABs for processing, perform upsampling operations, and then use GRU for deep fusion. Do this for features at $\frac{1}{8}$, $\frac{1}{4}$, $\frac{1}{2}$ and full resolution. Finally, a shallow convolutional decoder is used to output the features as defocus deblurred images $I_{DB} \in R^{H \times W \times 3}$.

2.5. Overall Loss Function

Here, we introduce the loss function used for training. Similar to previous work [1], we use L1-loss as the content loss L_c. Because L1-loss can recover high-frequency information more effectively. Compared to previous methods for defocus deblurring, we use a perceptual loss [16] L_p to update the model.

We use the losses above weighted to get L_G to train the model, where α is a hyperparameter to balance different types of loss.

$$L_G = L_c + \alpha \times L_p \tag{1}$$

3. EXPERIMENTAL RESULTS

3.1. Datasets

Like most methods [5, 3, 6], We use the dataset DPDD provided by [1] for training and testing. There are 500 sets of images in this dataset, and each set of images includes a defocused blurred image, a pair of DP views, and an all-in-focus (AiF) image, we follow its settings and divide 500 groups of pictures into 350, 74, and 76 groups according to the training set, validation set, and test set.

During the training phase, the hyperparameter α is set to 0.012, and the initial learning rate is set to 2×10^{-4}, which decreases by half every 30 epochs. and it gradually converges after 45 epochs. The batch size of both networks is set to 4 and optimized using the Adam optimizer, where $b1 = 0.9$, $b2 = 0.999$. We implemented the method using Pytorch and trained on an NVIDIA RTX 3090 GPU.

We also use the RealDOF test set [5] to verify the generalization of the network. The RealDOF test set contains 50 pairs of images, each pair consists of a defocused image and its corresponding all-in-focus image. We use the model trained on DPDD dataset to directly test on RealDOF test set.

3.2. Performance evaluation

Like many works, to evaluate the performance of defocus deblurring, we use the test set provided by [1] for testing. We compare the results with recent single-image defocus deblurring works. JNB [13], EBDB [2], and DMENet [14] are methods based on defocus maps. After they estimate the defocus map, they use non-blind deconvolution to defocus deblurring. DPDNet (single) [1], IFAN [5], MDPNet [6] and Defocus-GAN [7] are direct estimation methods that can directly restore AiF images.

For the above methods, we use the code and weights provided by the authors for testing. (IFAN uses data augmentation, we remove this method and retrain according to the code and training method provided by the authors.) For JNB, EBDB, and DMENet, following the advice of [1], we use the deconvolution method [17, 18] to recover the AiF image using the estimated defocus map. We also evaluate the number of network parameters in the inference stage to characterize the size of the model.

We use the commonly used metrics PSNR, SSIM, MAE, and LPIPS for defocus deblurring to evaluate the quality of the images. Table 2 shows the quantitative results of our method and other methods. Our method shows higher quality, outperforms all current methods with few model parameters, and restores image details to a great extent, improving the realism of images. Figure 3 shows a qualitative comparison. Traditional methods based on defocus maps and deconvolution have large blur areas. The effect of MDPNet and IFAN is greatly improved compared with the previous results, but often produces unnatural textures such as artifacts. For example, the texture of red walls and bronze figures. In particular, compared with these method, our method can better handle the texture of the image and recover the contours of objects. We can see that our method can better recover large-area blur, image details.

Method					Metrics		
Base	RCAB	GRU fusion	2-branch	3-branch	PSNR↑	SSIM↑	MAE↓
✓					24.77	0.766	0.042
✓	✓				25.26	0.775	0.041
✓	✓	✓			25.52	0.784	0.039
✓	✓	✓	✓		25.61	**0.789**	**0.038**
✓	✓	✓		✓	**25.65**	0.785	0.039

Table 3. Quantitative results of the ablation experiments on the DPDD dataset.

Fusion Method	PSNR↑	SSIM↑	MAE↓	LPIPS↓
Input fusion	25.41	0.780	0.040	0.116
Convolution fusion	25.23	0.782	0.040	0.114
GRU fusion	**25.52**	**0.784**	**0.039**	**0.113**

Table 4. Quantitative comparison of different deep fusion methods. Results are on the DPDD dataset.

3.3. Ablation study

Effects of each module. To demonstrate the effectiveness of each part of the module, we conduct ablation experiments in which all models are trained under the same conditions. Specifically, we use UNet [19] as a baseline model. Based on this, we gradually add RCAB and GRU components. We also gradually increase the parallel UNet-like branches and deeply integrate with GRU to better extract image information.

As can be seen in Table 3, RCAB can significantly improve the performance of deblurring due to the efficient fusion of channel information. But without the hint of depth information, its effect on large-scale blur is still not good. In the case of using GRU to fuse depth information, it performs better for the processing of letters with a greater degree of ambiguity. To further enhance the model's ability to deblur, we add a similar parallel branch structure, which brings better results.

The influence of different fusion methods. We want to know whether GRU has implemented an efficient integration method, so we compare three fusion methods, as shown in Table 4. One way is to feed the blurred image and the depth map together into the network for processing. One way is to use convolutional layers to deal with blurry image features and deep image features. One way is the GRU deep fusion method we proposed. The GRU fusion module performed the best. According to the mechanism of GRU, we think that the special updating and forgetting mechanism of GRU can effectively deblur with depth information.

4. CONCLUSION

We propose a single-image depth-enhanced defocus deblurring network. The proposed network can effectively handle large-area blur and effectively reconstruct image details and textures. In experiments, we verify the effect of each component in the model, achieving great performance with fewer parameters.

5. REFERENCES

[1] Abdullah Abuolaim and Michael S. Brown, "Defocus deblurring using dual-pixel data," in *Computer Vision – ECCV 2020*, Cham, 2020, pp. 111–126, Springer International Publishing.

[2] Ali Karaali and Claudio Rosito Jung, "Edge-based defocus blur estimation with adaptive scale selection," *IEEE Transactions on Image Processing (TIP)*, vol. 27, no. 3, pp. 1126–1137, 2018.

[3] Hyeongseok Son, Junyong Lee, Sunghyun Cho, and Seungyong Lee, "Single image defocus deblurring using kernel-sharing parallel atrous convolutions," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 2622–2630.

[4] Zhimin Lu, Jue Wang, Zhiwei Li, Song Chen, and Feng Wu, "A resource-efficient pipelined architecture for real-time semi-global stereo matching," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 2, pp. 660–673, 2022.

[5] Junyong Lee, Hyeongseok Son, Jaesung Rim, Sunghyun Cho, and Seungyong Lee, "Iterative filter adaptive network for single image defocus deblurring," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 2034–2042.

[6] Abdullah Abuolaim, Mahmoud Afifi, and Michael S. Brown, "Improving single-image defocus deblurring: How dual-pixel images help through multi-task learning," in *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2022, pp. 82–90.

[7] Jucai Zhai, Pengcheng Zeng, Chihao Ma, Jie Chen, and Yong Zhao, "Learnable blur kernel for single-image defocus deblurring in the wild," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 3, pp. 3384–3392, Jun. 2023.

[8] Qing Li, Jiasong Zhu, Jun Liu, Rui Cao, Qingquan Li, Sen Jia, and Guoping Qiu, "Deep Learning based Monocular Depth Prediction: Datasets, Methods and Applications," *arXiv e-prints*, p. arXiv:2011.04123, Nov. 2020.

[9] Juan L. Gonzalez and Munchurl Kim, "Forget about the lidar: Self-supervised depth estimators with med probability volumes," in *Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS)*, Red Hook, NY, USA, 2020, Curran Associates Inc.

[10] Clément Godard, Oisin Mac Aodha, and Gabriel J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6602–6611.

[11] Clement Godard, Oisin Mac Aodha, Michael Firman, and Gabriel Brostow, "Digging into self-supervised monocular depth estimation," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 3827–3837.

[12] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio, "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation," *arXiv e-prints*, p. arXiv:1406.1078, June 2014.

[13] Jianping Shi, Li Xu, and Jiaya Jia, "Just noticeable defocus blur detection and estimation," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 657–665.

[14] Junyong Lee, Sungkil Lee, Sunghyun Cho, and Seungyong Lee, "Deep defocus map estimation using domain adaptation," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 12214–12222.

[15] Lahav Lipson, Zachary Teed, and Jia Deng, "Raft-stereo: Multilevel recurrent field transforms for stereo matching," in *2021 International Conference on 3D Vision (3DV)*, 2021, pp. 218–227.

[16] Justin Johnson, Alexandre Alahi, and Li Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Computer Vision – ECCV 2016*, Cham, 2016, pp. 694–711, Springer International Publishing.

[17] David Fish, A. Brinicombe, Roy Pike, and J. Walker, "Blind deconvolution by means of the richardson–lucy algorithm," *JOSA A*, vol. 12, pp. 58–65, 01 1995.

[18] Dilip Krishnan and Rob Fergus, "Fast image deconvolution using hyper-laplacian priors," in *Proceedings of the 22nd International Conference on Neural Information Processing Systems (NeurIPS)*, Red Hook, NY, USA, 2009, p. 1033–1041, Curran Associates Inc.

[19] O. Ronneberger, P.Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. 2015, vol. 9351 of *LNCS*, pp. 234–241, Springer.