Defocus Image Deblurring Network With Defocus Map Estimation as Auxiliary Task

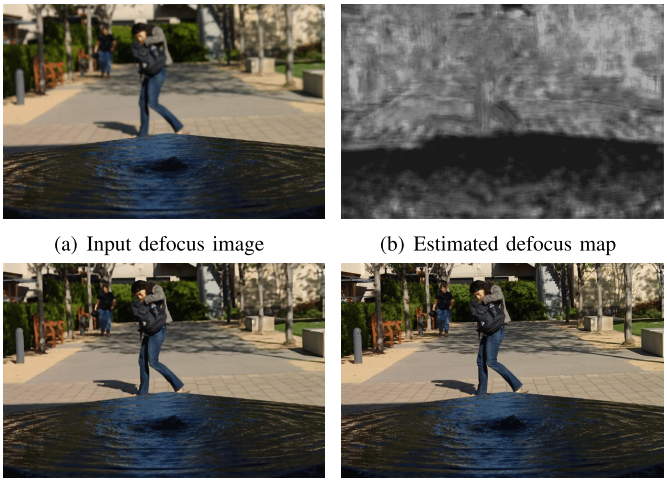
Haoyu Ma, Shaojun Liu, Qingmin Liao, *Senior Member, IEEE*, Juncheng Zhang, and Jing-Hao Xue, *Senior Member, IEEE*

***Abstract*— Different from the object motion blur, the defocus blur is caused by the limitation of the cameras' depth of field. The defocus amount can be characterized by the parameter of point spread function and thus forms a defocus map. In this paper, we propose a new network architecture called Defocus Image Deblurring Auxiliary Learning Net (DID-ANet), which is specifically designed for single image defocus deblurring by using defocus map estimation as auxiliary task to improve the deblurring result. To facilitate the training of the network, we build a novel and large-scale dataset for single image defocus deblurring, which contains the defocus images, the defocus maps and the all-sharp images. To the best of our knowledge, the new dataset is the first large-scale defocus deblurring dataset for training deep networks. Moreover, the experimental results demonstrate that the proposed DID-ANet outperforms the state-of-the-art methods for both tasks of defocus image deblurring and defocus map estimation, both quantitatively and qualitatively. The dataset, code, and model is available on GitHub: https://github.com/xytmhy/DID-ANet-Defocus-Deblurring.**

***Index Terms*— Defocus, deblurring, anxiliary learning, CNNs.**

I. Introduction

WHEN an image is captured, there are mainly two reasons for image blur, i.e., motion and defocus. On one hand, relative motion between the camera and the object, no matter which one is the actual moving one, leads to motion blur. On the other hand, when the depth range of the scene is relatively large but the depth of field of the camera is limited, the captured image can also be blurry due to defocus. An example of defocus blur is shown in Figure 1(a). The defocus amount is highly dependent on the depth between the object and the focal plane of the camera, therefore, it is usually spatially-varying. The defocus amount is usually modeled by the parameter of point spread function (PSF), forming a pixel-wise defocus map, as shown in Figure 1(b). Clear images can benefit many computer vision tasks such as detection, identification and segmentation [1], therefore, in this paper, we concentrate on defocus image deblurring from a single image.



(a) Input defocus image (b) Estimated defocus map

(c) Deblurring result (d) All-sharp ground truth

Fig. 1. Example for the proposed DID-ANet. The input defocus image and the ground truth all-sharp input come from the proposed dataset.

Defocus image deblurring from a single image (Figure 1(c)) is a challenging ill-posed problem [2], as the target all-sharp image (Figure 1(d)) contains much more details than the input defocus image (Figure 1(a)), especially in the highly defocused areas such as the background including the people and chair in Figure 1. Hence more attention should be paid to the highly defocused areas by incorporating external information. Fortunately, by using deep neural networks, rich real-world information could be learned from the data from a wide range of environments. Therefore, we propose a deep network approach for defocus image deblurring. Our contributions are four-fold.

1) We design a new network for defocus image deblurring, with defocus map estimation as auxiliary learning task. The proposed **D**efocus **I**mage **D**eblurring **A**uxiliary Learning **Net**work (**DID-ANet**) is a novel deep learning (end-to-end) architecture specifically designed for defocus deblurring.

2) We introduce a new defocus image deblurring dataset for training and test of deep networks. This dataset contains

Manuscript received June 30, 2021; revised September 20, 2021 and October 31, 2021; accepted November 2, 2021. Date of publication November 18, 2021; date of current version December 8, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 61771276 and in part by the National Key Research and Development Program of China under Grant 2016YFB0101001. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Emanuele Salerno. *(Haoyu Ma and Shaojun Liu contributed equally to this work.) (Corresponding author: Qingmin Liao.)*

Haoyu Ma, Qingmin Liao, and Juncheng Zhang are with the Division of Information Science and Technology, Tsinghua Shenzhen International Graduate School, Shenzhen 518055, China, and also with the Department of Electronic Engineering, Tsinghua University, Beijing 100084, China (e-mail: liaoqm@tsinghua.edu.cn).

Shaojun Liu was with the Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong SAR, China. He is now with the College of Health Science and Environmental Engineering, Shenzhen Technology University, Shenzhen 518118, China.

Jing-Hao Xue is with the Department of Statistical Science, University College London, London WC1E 6BT, U.K.

Digital Object Identifier 10.1109/TIP.2021.3127850

partial defocus images, all-sharp (i.e., all-in-focus) images, as well as the corresponding defocus maps. The partial defocus image and all-sharp image are generated from a single image captured by a light-field camera. As far as we know, this is the *first* large-scale defocus blurring dataset taken in real scenes which could be used for the training of deep learning networks.

3) By using defocus map estimation as guidance for defocus image deblurring, we alleviate the difficulty in network training. Furthermore, we improve the deblurring results by introducing effective loss functions and flexible training strategies.

4) Our experiments on several benchmarks show that DID-ANet obtains the state-of-the-art performance on both the defocus map estimation and defocus image deblurring, both quantitatively and qualitatively.

II. Related Work

A. Image Deblurring

For defocus image deblurring, a typical blur model can be expressed as follows [3]:

$$I_{Blur} = k * I_{Clear} + N_G, \tag{1}$$

where I_{Clear} is the clear image, k is the blur kernel, N_G is the additive Gaussian noise, and I_{Blur} is the blurry image [4]. Usually, the defocus procedure is modeled as a convolution of the clear image with the PSF; therefore, conventional methods [5]–[7] first estimate defocus kernels and then deconvolve the defocus image to produce the all-sharp image. This is a direct way and can usually obtain satisfactory results for areas with low or middle level defocus, but cannot produce sharp results for areas with high level defocus since almost all the high frequency information has been lost. Additional natural images priors [8]–[10] can improve the deblurring results in several particular scenes, but they are insufficient to fill in the rich real-world information, either. Moreover, due to the scene change in photos taken in different environments, it is usually difficult to obtain the natural scene priors with conventional methods.

In recent years, with the development of deep learning methods, many new approaches have been proposed. In the early stage, Sun *et al.* [11] estimate the blur kernel with CNN; Chakrabarti [12], Anwar *et al.* [13], [14] and Gong *et al.* [15] also use neural networks to replace several parts of the deblurring process. Currently, end-to-end CNN approaches are widely applied. Nah *et al.* [16] and Tao *et al.* [4] use multi-scale structure [17] for dynamic scene deblurring and get good visual results. Meanwhile, Ramakrishnan *et al.* [18] and Kupyn *et al.* [3], [19] use the conditional adversarial networks (GANs) [20] for the deblurring. Unfortunately, most of these end-to-end methods pay no attention to different types of blur, and some are specially designed for the motion blur and therefore unsuitable for the defocus deblurring.

Recently, Abuolaim *et al.* [21] use CNN to deblur defocus dual-pixel image pairs. But for single image defocus deblurring, their results are unsatisfactory. Moreover, Lee *et al.* [22] apply CNN for defocus image deblurring with the dual-pixel image dataset [21]. In this paper, we want to introduce a novel model and a new light-field camera dataset for single defocus image deblurring.

B. Defocus Map Estimation

Existing defocus map estimation (DME) methods can be roughly categorized into three classes, i.e., edge-based methods, region-based methods and learning-based methods.

In the edge-based methods, the defocus amount is usually calculated at edge points and then propagated to the whole image. Zhuo and Sim [23], Cao *et al.* [24], Zhang *et al.* [25] and Karaali and Jung [26] reblur the input image with Gaussian kernels and use the ratio of the gradients of the reblurred images at edge points to calculate the defocus amount. Liu *et al.* [27] propose a two-parameter defocus model to better analyze the defocus process and produce more accurate estimations at edge points. Park *et al.* [28] unify handcrafted and deep features to estimate defocus amount at edge points. After the defocus amounts are obtained at edge points, propagation method such as Laplacian matting, KNN matting or guided image filtering is employed to obtain the final defocus map. In this procedure, the input image or a smoothed version is used as the guidance. Therefore, the final defocus map usually suffers from the textures of the input image. Moreover, for areas that are far from edge points, the propagated defocus amount estimations are usually not reliable enough.

For region-based methods, the defocus amounts are often directly calculated from local patches centered at the current pixel. Trouvé *et al.* [29] deblur the input image patch with a set of PSF candidate and take the one which produces the sharpest deblurred result as the estimation of defocus amount. Shi *et al.* [30] build a defocus patch dictionary on which they decompose the input image patches and use the sparsity of the decomposition coefficient as a feature. Zhu *et al.* [31] employ localized 2D frequency analysis to generate the likelihood of defocus amount and employ coherent labeling to refine it. D'Andrès *et al.* [5] extract a feature from the likelihood and refine it with a regression tree fields to ease the problem of [31]. They also build a realistic dataset for DME in the sense of image deblurring, using a light field camera. This is the first spatially-varing DME dataset, however, there are only 22 images. Usually, region-based method is free from textures of the input image, while they often produce inaccurate estimations for homogeneous areas and cannot catch the defocus discontinuities very well. In our another work [32], we extract a region-based feature based on improved likelihood and incorporate it with edge-based basis to produce texture-free defocus map while catching the defocus discontinuities well.

Recently, there are several deep-learning-based trials. Yan and Shao [33] build a general regression neural network to first classify the blur type and then estimate the blur parameter. The defocus amount of their training dataset is spatially invariant, limiting the application of their method. Zhao *et al.* [34] use a bottom-top-bottom fully convolutional network to detect defocus blur, which can be viewed as a loose problem of DME. Similarly, Zhang *et al.* [35] build a smart defocus dataset with three defocus levels and train a deep neural network to estimate the defocus of an input image. Lee *et al.* [36] build a synthetic dataset based on which they

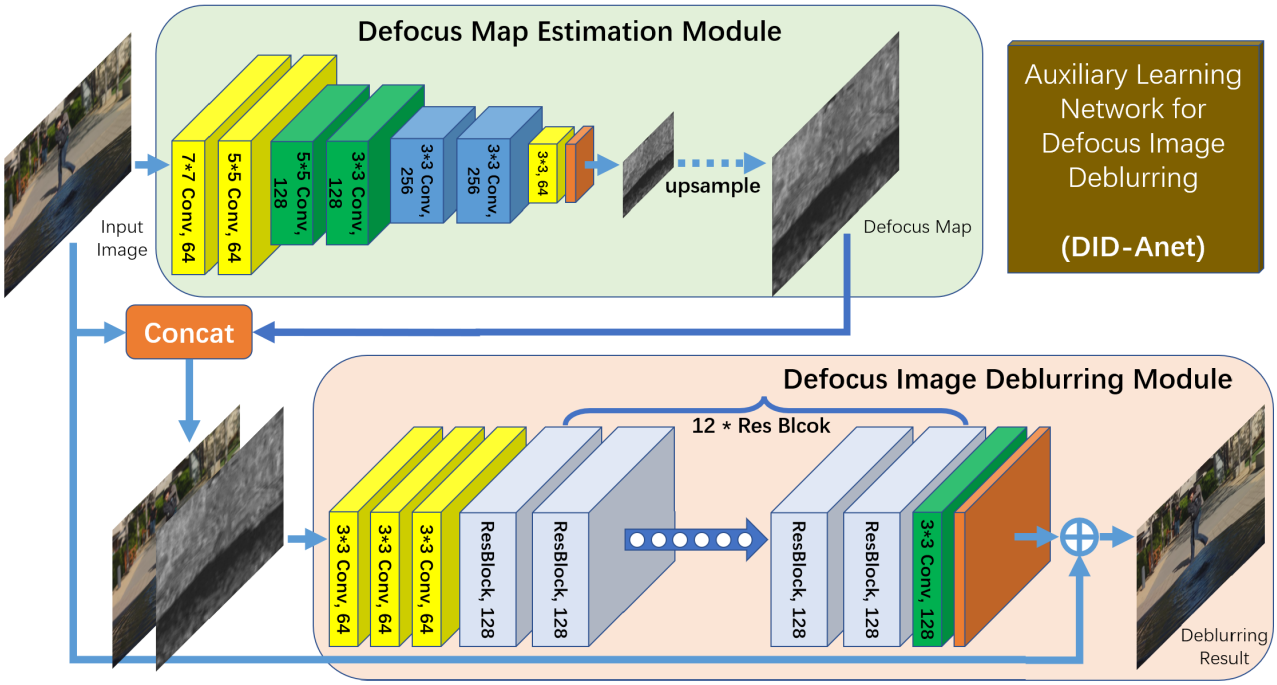


Fig. 2. The architecture of proposed DID-ANet. The network consists of two main parts: the defocus map estimation sub-net and the defocus image deblurring sub-net. The image is deblurred with the guidance of the estimated defocus map.

develop an end-to-end deep neural network (DME-Net) to generate defocus map. Domain adaptation is used since there are not enough data with ground truth defocus map. This is the first truly deep-learning based DME method, while the lack of real scene data limits its performance and applications.

C. Auxiliary Learning

Auxiliary learning is a method to complement the primary task by training on additional auxiliary tasks alongside this primary task [37]. A direct approach to auxiliary learning is to use a related task as auxiliary. Intermediate representations are used as auxiliary supervision at lower levels of deep networks to combine the advantages of end-to-end training and more traditional pipeline approaches [38], [39]. Liebel and Körner [40] empirically demonstrate that auxiliary tasks can boost network performance, in terms of both final results and training time. Several different vision auxiliary tasks have been applied for depth estimation in monocular or multiple images [41]–[43]. Jaderberg *et al.* [44] use unsupervised learning tasks to continue developing in the absence of extrinsic rewards in reinforcement learning.

In this paper, we use defocus map estimation as the auxiliary task due to its close relevance to the primary task: defocus image deblurring. The network architecture is also designed according to the relationship between these two tasks. In short, defocus map estimation is used as a low-level guidance in front of the defocus image deblurring.

III. Defocus Image Deblurring

A. Auxiliary Learning Network

As there are out-of-focus images and clear ground truth in pair, it is a natural thought to train a single end-to-end network solving the defocus image deblurring. However, our experience with such a simple structure is that the network output is very similar to the original input, and the end point error (EPE) would not decrease to a desired small value. One explanation for this phenomenon is: since the input and output are very similar in large scale and there are clear areas in the input images, the output similar to the input can be a trivial local minimum, where the clear areas reach the smallest loss and the out-of-focus areas gets a reasonable EPE, making it hard to be further enhanced.

To avoid such a local minimum, we can group the input pixels according to their defocus amount, and the pixels with similar defocus amount can be deblurred with the same network parameters. That is, the defocus map can be used to guide the deblurring process. Hence, we propose in this paper a defocus image deblurring network with defocus map estimation as auxiliary task. As shown by the detailed architecture in Figure 2, the input defocus image is firstly processed by a simple defocus map estimation sub-net, which contains several convolution layers to estimate a defocus map 4 times smaller than the input image; then the estimated defocus map is upsampled to the original size and concatenated with the input images as the input of the defocus deblurring sub-net, which contains 12 res blocks and several convolution layers. With the defocus map as the guidance for deblurring, the defocus areas in the input image are deblurred nicely. The final deblurring result is the deblurring residual added to the original input image.

The standard residual blocks in our network contain 2 convolutional layers, 2 batch normalization layers, and a ReLU layer in the middle [45]. The kernel size is 128 in the defocus deblurring sub-net. No pooling or sub-sampling is used in the defocus deblurring sub-net as [46].

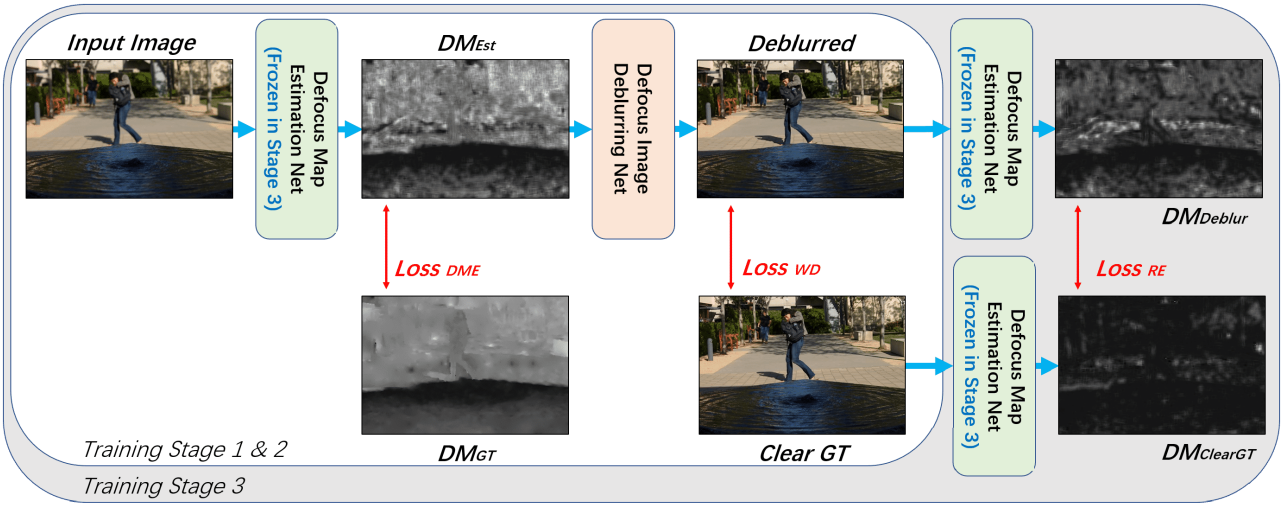


Fig. 3. Illustration for the estimation and re-evaluation of the defocus map, as well as the corresponding training losses employed in our DID-ANet. In the first two training stages shown above, the two sub-nets are optimized jointly. In the last training stage, the defocus map estimation sub-net is frozen, focusing on the improvement of deblurring sub-net.

In addition, our network will pay more attention to the areas with larger defocus, as to be introduced in detail in Section III-C of loss functions.

B. Defocus Re-Evaluation

The ultimate goal of the defocus image deblurring is to generate an all-in-focus image. It is natural to propose that the defocus map estimation sub-net can be reused to evaluate the deblurring effect, as shown in Figure 3. Therefore, to further enhance the deblurring result after the first round of training, a re-evaluation loss is applied: the deblurring result and the ground truth clear image are separately processed by the frozen defocus map estimation sub-net, and then the average pixel difference between these two new estimated defocus maps is used as the re-evaluation loss. If the deblurring result is similar to the ground truth, the defocus map estimations of the deblurring result and the ground truth would be similar, too.

C. Loss Functions

To train the proposed network, we elaborately adopt several kinds of loss functions. The commonly used L_1 norm and L_2 norm are firstly applied. We use the defocus map estimation loss ($Loss_{DME}$) to supervise the estimated defocus map (DM_{Est}) with the ground truth defocus map (DM_{GT}):

$$Loss_{DME} = \|DM_{Est} - DM_{GT}\|_1. \tag{2}$$

Then, we design a loss to supervise the deblurring result with its paired clear ground truth. Because highly defocused areas are much more difficult to deblur than lightly defocused or non-defocused areas, with the estimated defocus map as the reference, we increase the importance of the difficult areas by using the weighted deblur loss ($Loss_{WD}$) with different weights at different positions:

$$Loss_{WD} = \|W_{DME} \times (Deblurred - ClearGT)\|^2, \tag{3}$$

where weight map W_{DME} is the normalized defocus map with an offset W_0:

$$W_{DME} = \frac{DM_{Est}}{mean(DM_{Est})} + W_0. \tag{4}$$

Here we use $W_0 = 1/9$.

As mentioned in Section III-B, we reuse the defocus map estimation sub-net to evaluate the deblurring result of the defocus image deblurring network to enhance it. That is, we compare the defocus map estimations of the deblurring output ($DM_{Deblurred}$) and the all-sharp ground truth ($DM_{Clear_{GT}}$). The difference is called the re-evalution loss $Loss_{RE}$:

$$Loss_{RE} = \|DM_{Deblur} - DM_{ClearGT}\|_1. \tag{5}$$

Accordingly, we also design some training strategies to optimize the whole network. Specifically, the training procedure contains three stages (Figure 3). In stage 1 and stage 2, the two sub-nets are jointly trained for 400 and 200 epochs, respectively. For stage 1, we employ the ground truth defocus map as the input of the defocus image deblurring sub-net to avoid divergence caused by random output of the defocus map estimation sub-net and speed up the training. While for stage 2, we use the output of the defocus map estimation sub-net as the input of the defocus image deblurring sub-net to jointly fine-tune the whole network. In stage 1 and stage 2, we employ $Loss_{DME}$ and $Loss_{WD}$ for supervision:

$$Loss_1 = \lambda_1 \times Loss_{DME} + \lambda_2 \times Loss_{WD}. \tag{6}$$

In stage 3, we add the re-evaluation loss to further fine-tune the defocus image deblurring sub-net for another 400 epochs, with parameters of the defocus map estimation sub-net frozen. Hence we use $Loss_{WD}$ and $Loss_{RE}$ for supervision:

$$Loss_2 = \lambda_2 \times Loss_{WD} + \lambda_3 \times Loss_{RE}. \tag{7}$$

In this paper, the weights of the loss functions are $\lambda_1 = 0.1$, $\lambda_2 = 0.9$ and $\lambda_3 = 0.2$.

D. DED Real Scenes Dataset

To the best of our knowledge, in the sense of defocus map estimation and defocus image deblurring, there is only a small dataset called Realistic [5] consisting of 22 image pairs, which are far from enough for training deep neural networks. To fill this gap and facilitate the training of our model, we build the first large-scale realistic dataset for defocus map estimation and defocus image deblurring (termed as DED dataset) with a light field camera.

Usually, it is extremely hard to directly capture an RGB-Defocus dataset using conventional cameras. To build such a dataset, typically two images captured with different camera settings are needed. However, the contents and intensities of these two images would be different more or less. Consequently, geometric and photometric alignments are needed. Unfortunately, precise alignments are also difficult. Alternatively, one can estimate defocus maps based on stereo/RGB-Depth datasets and then reblur the all-in-focus images manually to synthetically generate the partially defocus images. However, the employed kernels might be different from real ones and consequently the produced defocused images are different from real scenes. To bypass these problems, we use a Lytro Illum light field camera [47], which can generate two differently focused images at one shot, to generate the dataset.

The Lytro company provides a software along with their camera to process the captured images that record the 4 dimensional light field. With the help of this software, the all-sharp image I_s, a partially-defocus image I_b and the corresponding depth map I_d[1] can be easily generated. In principle, I_s and I_b are generated by filtering the 4 dimensional light field with specific 4 dimensional band-pass filters [48]. They can be viewed as if they were captured by a camera twice with different settings, based on the Fourier Slice Photography Theorem [49] for light field camera. I_d is generated using the stereo information extracted from the 4 dimensional light field.

Then, inspired by [5], we calculate the mean squared error (MSE) between I_b and a reblurred version of I_s in a patch-wise way as follows:

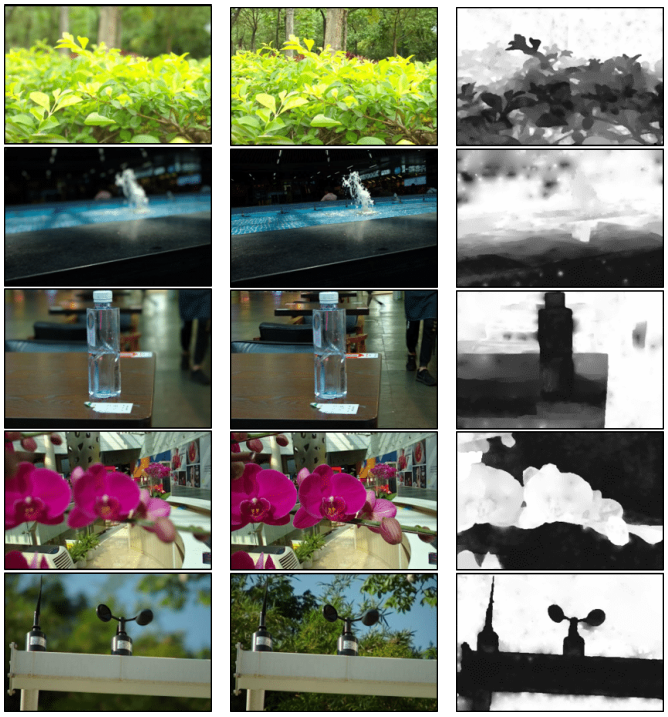
$$d(r)[i] = \frac{\sum_{j \in \mathcal{N}_i} (I_b[j] - (I_s \otimes k(r))[j])^2}{L^2}, \tag{8}$$

where i is a pixel, $\mathcal{N}_i$ is a small window of size $L \times L$ centered at pixel i, and r is the radius of the candidate PSF $k(r)$. Then the defocus amount at pixel i is obtained by minimizing this MSE:

$$b[i] = r^* = \min_r d(r)[i]. \tag{9}$$

Next, we detect the high confidence values as [5] did and propagate them to the non-confident pixels via Laplacian matting with the depth map I_d as the guidance.

[1]Please note that for a Lytro Illum, the maximum spatial resolution is 625 × 433 [48] while the resolution of the Lytro software output is 2450 × 1634. Therefore, we down-scale the output of the Lytro software with a factor of 0.25, i.e., the resolution of I_s, I_b and I_d is 613 × 409. The down-scaling also helps keep I_s sharp enough to serve as the all-sharp ground truth.



(a) Defocus image (b) All-sharp image (c) Defocus map

Fig. 4. Examples of the proposed DED dataset.

In the end, we generate in total 1,112 image pairs in the proposed DED dataset and Figure 4 illustrates several examples. Some of the images are from the multi-view dataset [50] and the others (over a half) are captured by ourselves. Among the image pairs, 100 pairs are randomly selected as the test set, and the rest 1,012 pairs are for the training. The selection is also conducted to ensure that the test set has different scenes from the training set. Both the training set (including the defocus images, the defocus maps and the clear ground truth) and the test sets (only the defocus images) of DED are open source.

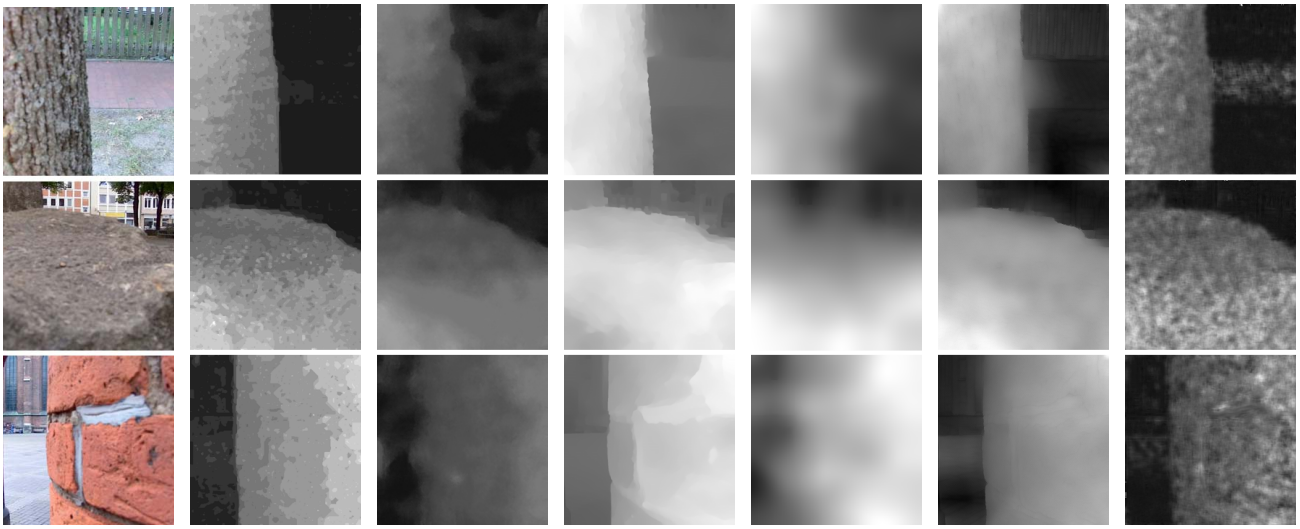
IV. Experiments

A. Implementation Details

For the training process, the Adam solver is used with parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 10^{-8}$ with the numbers of epochs detailed above. The input images and the corresponding all-sharp ground truths, as well as the defocus maps are randomly cropped to the size of 256 × 256. Other data augmentation strategies, such as random flipping, rotation and color change, are also applied to make the dataset more variable [51]. The batch size is set to 16 when using 4 × Nvidia 1080Ti GPU for training, and the testing is conducted with a single GPU. The testing time for a single image of size 600 × 400 is 0.27 seconds on average, with about 570 billion FLOPs and 19.2 million parameters.

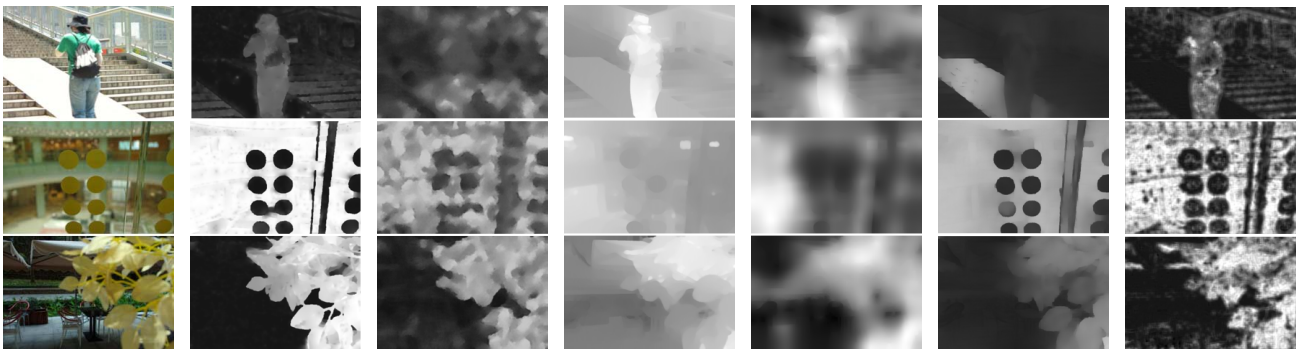
B. Experimental Results

We evaluate the proposed method on the Realistic dataset [5] and the test set of the proposed DED dataset (DED-test). Both the results of defocus map estimation and



(a) Input images (b) Ground truth (c) D'Andrès [5] (d) Park [28] (e) Karaali [26] (f) DME-Net [36] (g) Ours

Fig. 5. Visual comparison of defocus map estimation on realistic.



(a) Input image (b) Ground truth (c) D'Andrès [5] (d) Park [28] (e) Karaali [26] (f) DME-Net [36] (g) Ours

Fig. 6. Visual comparison of defocus map estimation on DED-test.

TABLE I

QUANTITATIVE COMPARISON FOR DEFOCUS MAP ESTIMATION, WHERE THE BEST RESULTS ARE IN BOLD

Method	D'Andrès *et al.* [5]	Park *et al.* [28]	Karaali *et al.* [26]	DME-Net [36]	Ours
MAE, Realistic	**0.1968**	0.2881	0.3547	0.3105	0.2392
MSE, Realistic	**0.0941**	0.1728	0.2200	0.1727	0.0985
MAE, DED-test	0.3321	0.3423	0.3931	0.3863	**0.2443**
MSE, DED-test	0.1198	0.1216	0.1627	0.1538	**0.0644**

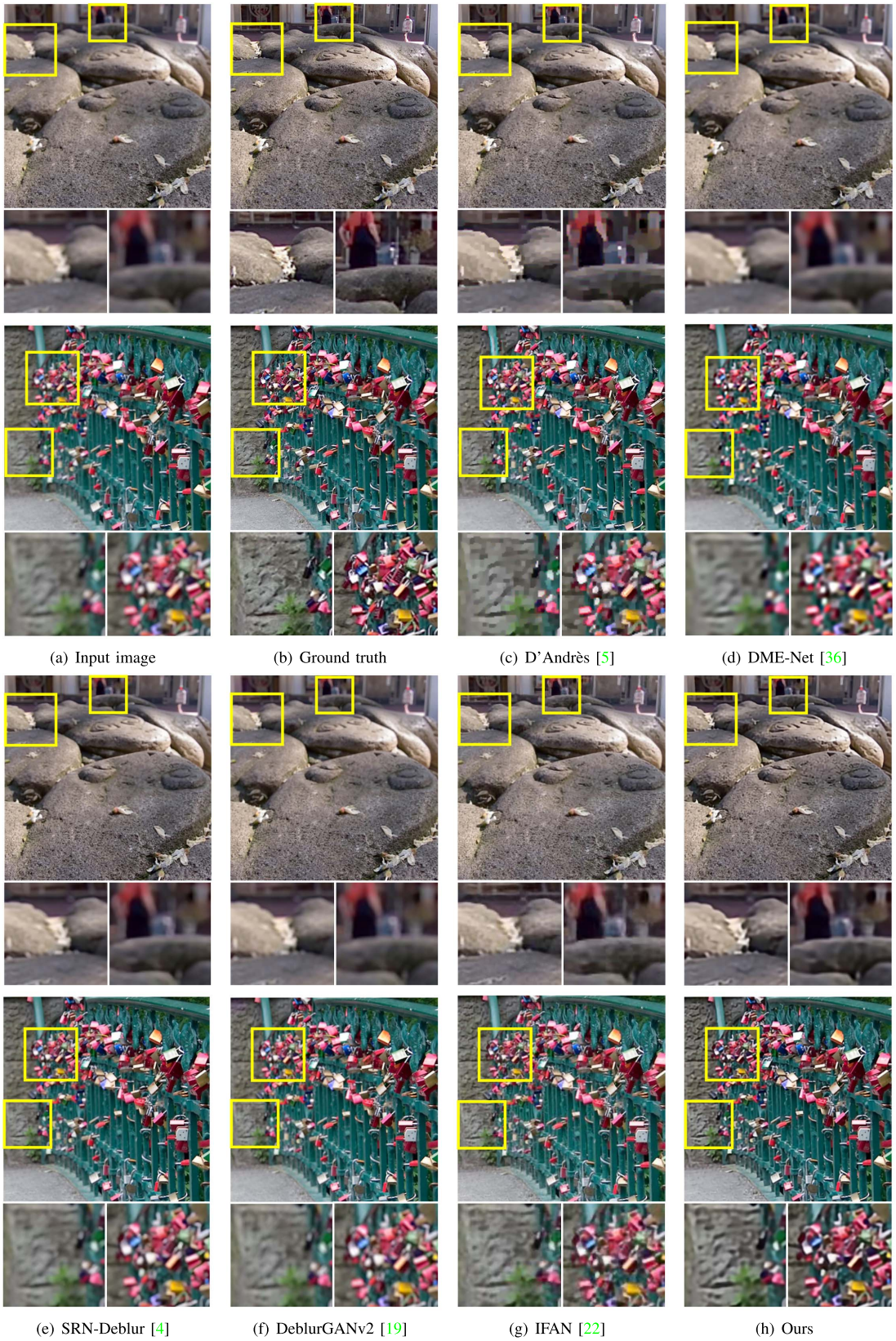
defocus image deblurring are compared with the state-of-the-art methods.

The results for defocus map estimation are compared with the methods of Zhou and Sim [23], D'Andrès *et al.* [5], Park *et al.* [28], Karaali and Jung [26] and the recent deep-learning-based DME-Net [36]. The evaluation metrics are the mean absolute error (MAE) and mean squared error (MSE) to the defocus map ground truth. The quantitative comparison can be found in Table I, where the best results are in bold. We can see that: for Realistic, the proposed method is comparable with D'Andrès *et al.* [5] and outperforms the rest three methods; for DED-test, the proposed method performs the best, with the lowest MAE and MSE.

Several visual examples of defocus map estimation are also shown in Figure 5 (Realistic) and Figure 6 (DED-test). Our results are much closer to the ground truth and the error area beyond the boundary is less than those of other four methods.

The results for defocus deblurring are compared with the conventional method of D'Andrès *et al.* [5]; the deep learning methods of SRN-Deblur [4], DeblurGAN [3], [19], IFAN [22]; and the DME-Net [36] which applies CNN for defocus map estimation and conventional deconvolution [52] for deblurring. The deep learning methods are fine-tuned[2] on the training set of the proposed DED dataset. The evaluation metrics are the Peak Signal to Noise Ratio (PSNR) and Structural Similarity (SSIM) to the clear ground truth. The quantitative comparison can be found in Table II, where the best results are in bold. For both Realistic dataset and the proposed DED test set, the proposed method outperforms all the other methods; with the best PSNR and SSIM.

[2]The learning rates are set separately according to the original papers, and the training processes take 200 epochs, to make sure the models are convergent.

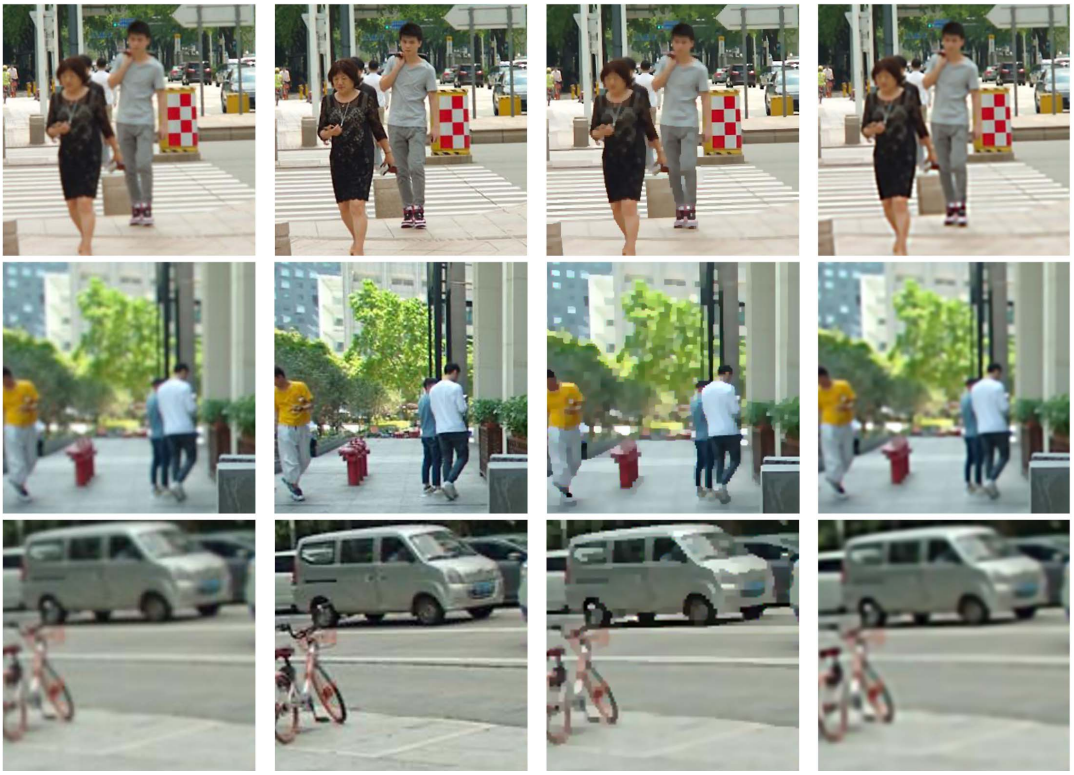


(a) Input image (b) Ground truth (c) D'Andrès [5] (d) DME-Net [36]

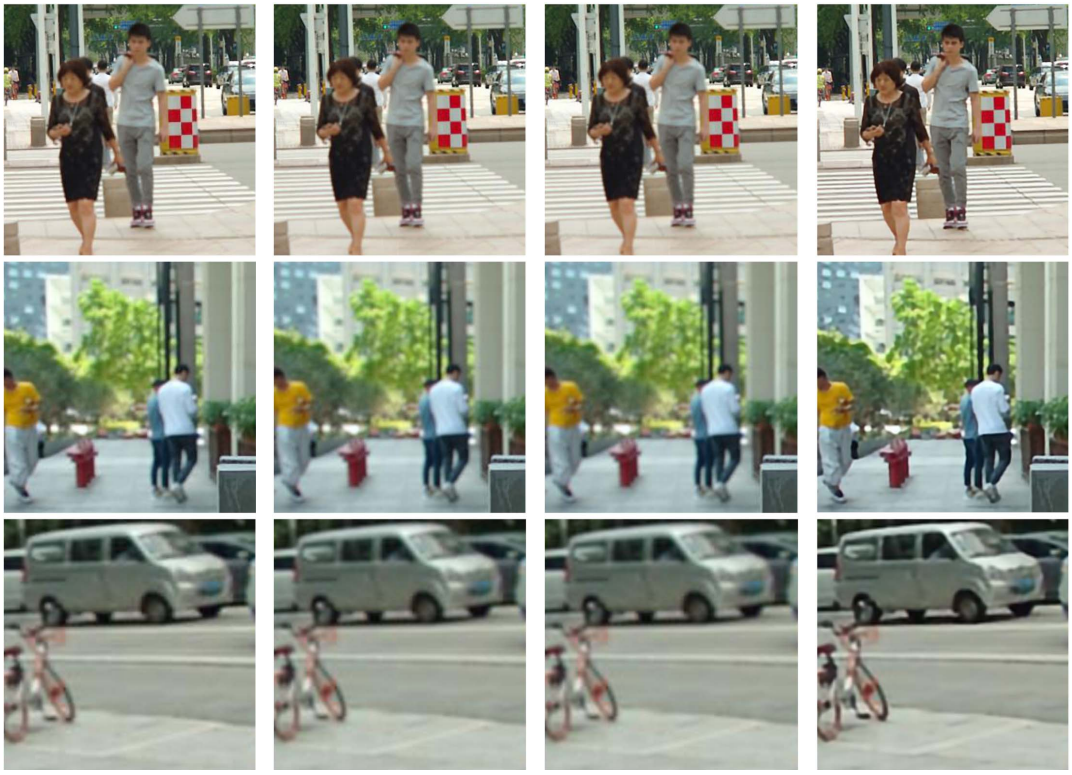
(e) SRN-Deblur [4] (f) DeblurGANv2 [19] (g) IFAN [22] (h) Ours

Fig. 7. Visual comparison of defocus image deblurring on realistic.

Several visual examples of defocus image deblurring are shown in Figure 7 (Realistic) and Figure 8 (DED-test). As images from the DED dataset are larger, we crop them and zoom in for a better view. Results of the proposed DID-ANet are more clear and vivid in our results than those of all other methods. In Figure 7 (top), the texture in the rock



(a) Input image (b) Ground truth (c) D'Andrès [5] (d) DME-Net [36]



(e) SRN-Deblur [4] (f) DeblurGANv2 [19] (g) IFAN [22] (h) Ours

Fig. 8. Visual comparison of defocus image deblurring on DED-test. As the images in DED dataset are too large, we crop them and zoom in for better view.

and the people with a black bag are clearer in our result; in Figure 7 (bottom), our result has a much more colorful number '2', and more vivid boundaries of the locks. In Figure 8, the human faces (top), the pedestrians and the background trees (middle), and the cars and the bicycle (bottom) are all more realistic and clearer in our result. Besides, the visually fuzzy phenomenon is greatly decreased in our deblurring results.

TABLE II

QUANTITATIVE COMPARISON FOR DEFOCUS IMAGE DEBLURRING, WHERE THE BEST RESULTS ARE IN BOLD

Method	D'Andrès *et al.* [5]	DMENet [36]	SRN-Deblur [4]	DeblurGANv2 [19]	IFAN [22]	Ours
PSNR, Realistic	25.4268	24.0397	24.2156	24.3922	24.5410	**26.0803**
SSIM, Realistic	0.8504	0.7615	0.7948	0.7942	0.8234	**0.8559**
PSNR, DED-test	27.7139	26.0376	28.1028	28.6837	27.9823	**31.0091**
SSIM, DED-test	0.9146	0.8719	0.9261	0.9305	0.9311	**0.9533**

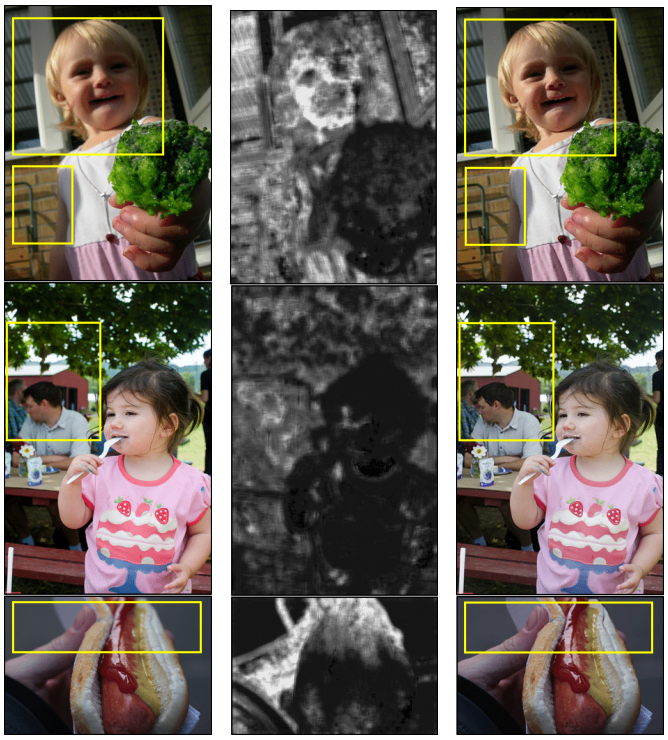
TABLE III

ABLATION STUDY ON REALISTIC AND DED TEST SETS., WHERE THE BEST RESULTS ARE IN BOLD. THE DEBLURRING RESULTS CAN BENEFIT FROM THE AUXILIARY DEFOCUS MAP ESTIMATION TASK, THE LOSS FUNCTIONS AND FLEXIBLE TRAINING STRATEGIES

Method	DME+Deconv [52]	Backbone	Without AL	Stage 1, L_2	Stage 1, L_{WD}	Stage 2	Stage 3
PSNR, Realistic	23.2219	25.6590	25.6680	25.9074	25.9158	25.9288	**26.0803**
SSIM, Realistic	0.7448	0.8431	0.8403	0.8474	0.8507	0.8514	**0.8559**
PSNR, DED-test	27.6792	30.5059	30.6080	30.6590	30.6867	31.0008	**31.0091**
SSIM, DED-test	0.9158	0.9466	0.9469	0.9475	0.9478	0.9503	**0.9533**

In contrast, there is block effect in results of [5], and the methods cannot deal with the defocus blur very well. Specifically, for the DME-Net [22] with deconvolution deblurring [52], the results are not as clear as ours; for the motion deblurring methods [4], [19], they cannot deal with the defocus blur very well; and for the IFAN [22], although it is specially designed for defocus blur removal and the model is refined with the proposed DED dataset, the results are still not as clear as those of the proposed method. Moreover, there are noticeable artifacts in the deblurring results of [22], which also lead to lower PSNR and SSIM scores than our method, as shown in Table II.

Moreover, to verify the generalization ability of the DID-ANet besides the light field camera generated datasets, we select a small collection of pictures with obvious defocus areas from the COCO dataset (some examples in Figure 9). The higher value in the defocus map, the higher defocus amount. After deblurring, defocused areas (the face of the girl and the wall in the first example, the man sitting behind the desk in the second example, as well as the above part in the third example) become clear.



(a) Input image (b) Defocus map (c) Deblurring results

Fig. 9. Examples of the defocus map estimation and defocus deblurring by DID-ANet on the COCO dataset.

C. Ablation Studies

Several ablation studies are conducted on both the Realistic dataset and the test set of the DED dataset. The results can be found in Table III.

Firstly, to demonstrate the necessity of auxiliary learning, two experiments are conducted. One is termed "Backbone", the simple structure without the defocus map estimation module. The other is termed "Without AL", the complete network as DID-ANet but trained with no supervision for defocus map estimation. As expected, these two variations produce much lower PSNR and SSIM for defocus image deblurring. Furthermore, more training epochs are required for both "Backbone" and "Without AL" to convergence, meaning that it is hard to train a network for defocus image deblurring without the auxiliary learning. Besides, we also use the deconvolution method [52] to deblur the input image with the defocus map generated by our DME module. However, the PSNR and SSIM of the results are much lower than even the "Backbone" network. This implies that the network methods are indeed important for defocus image deblurring.

Then, another ablation experiment is conducted to verify the effectiveness of each training strategy. Specifically, we test the deblurring result after each training stage. As shown in Table III, the performance is already better than those of all other methods after the first training stage (compared with Table II). Moreover, after each training stage, both PSNR and SSIM get increased. We also study the effectiveness of each loss function in the proposed DID-ANet. The performance with a simple L_2 loss is worse than that with the weight loss (L_{WD}), meaning that L_{WD} is useful. Training stage 3 can improve the performance compared with training stage 2, showing the effectiveness of L_{RE}.

We also show the loss curves for the training processes of the different network structures or loss functions aforementioned in Figure 10. Only the first 200 training epochs are

TABLE IV

FIVE-FOLD CROSS VALIDATION OF DID-ANET ON THE PROPOSED DED DATASET. THE MODEL IS TRAINED ON FOUR FOLDS AND TESTED ON THE REMAINING FOLD AS WELL AS THE REALISTIC DATASET. IT SHOULD BE NOTED THAT THE PERFORMANCE DOES NOT CHANGE VERY MUCH FOR THE REALISTIC DATASET, MEANING THAT THE MODEL IS INSENSITIVE TO THE TRAINING SET PARTITION

Metric, tested on \ Model trained on	Folds #2-5	Folds #1, 3-5	Folds #1, 2, 4, 5	Folds #1-3, 5	Folds #1-4
PSNR, Defocus Map Realistic	25.9831	26.0803	26.0467	26.0056	26.0389
SSIM, Realistic	0.8556	0.8559	0.8540	0.8549	0.8523
PSNR, DED remaining fold	30.2642	28.4579	29.6008	28.5884	26.7559
SSIM, DED remaining fold	0.9384	0.9139	0.9308	0.9165	0.8966

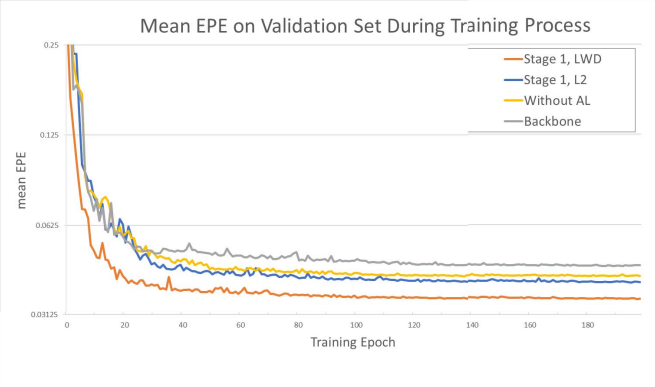


Fig. 10. EPE loss curve on validation set for ablation studies. The proposed DID-ANet with the weighted deblur loss function L_{WD} achieves the lowest EPE and can converge more quickly than all other variations.

plotted in the figure. We use the mean end point error (EPE) on the validation set (about 10% of the training set that are kept out for validation in the training process) to show the performance of all the variations. As shown in Figure 10, the proposed DID-ANet with the weighted deblur loss function L_{WD} achieves the lowest EPE and can converge more quickly than all other variations.

D. Cross Validation on the DED Dataset

A five-fold cross validation experiment is conducted on the proposed DED dataset to demonstrate the robustness of the proposed DID-ANet. The DED dataset is partitioned to 5 folds randomly. Specifically, we first divide apart all the images according to the different scenes, where each scene contains 2 to 6 images. Then, the scenes are randomly assigned to a fold. Finally, the 5 folds are adjusted appropriately to make sure that each fold has about 20% images.

The DID-ANet is trained on 4 folds and tested on the remaining fold. Furthermore, the models are also tested on the Realistic dataset. The results are shown in Table IV. It should be noted that the performance for these models are quite similar to each other for the Realistic dataset, meaning that the proposed model is insensitive to the training/test partition.

V. CONCLUSION

In this paper, we propose a novel deep auxiliary learning approach called DID-ANet, with defocus map estimation as the auxiliary task for defocus image deblurring. The guidance provided by the defocus map estimation makes the network easier to train end-to-end, and helps to improve deblurring results. Several novel loss functions and flexible training strategies are also introduced. Furthermore, a new large-scale defocus dataset termed DED is built, which is also the first large-scale defocus deblurring dataset taken in real scenes and suitable for training deep networks. Experiments show that our DID-ANet obtain the state-of-the-art performance for both defocus map estimation and defocus image deblurring tasks, both quantitatively and qualitatively.

REFERENCES

[1] R. Hassen, Z. Wang, and M. M. A. Salama, "Objective quality assessment for multiexposure multifocus image fusion," *IEEE Trans. Image Process.*, vol. 24, no. 9, pp. 2712–2724, Sep. 2015.
[2] P. Campisi and K. Egiazarian, *Blind Image Deconvolution: Theory and Applications*. Boca Raton, FL, USA: CRC Press, 2016.
[3] O. Kupyn, V. Budzan, M. Mykhailych, D. Mishkin, and J. Matas, "DeblurGAN: Blind motion deblurring using conditional adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8183–8192.
[4] X. Tao, H. Gao, X. Shen, J. Wang, and J. Jia, "Scale-recurrent network for deep image deblurring," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8174–8182.
[5] L. D'Andrès, J. Salvador, A. Kochale, and S. Süsstrunk, "Non-parametric blur map regression for depth of field extension," *IEEE Trans. Image Process.*, vol. 25, no. 4, pp. 1660–1673, Apr. 2016.
[6] A. K. Katsaggelos and K. T. Lay, "Maximum likelihood blur identification and image restoration using the EM algorithm," *IEEE Trans. Signal Process.*, vol. 39, no. 3, pp. 729–733, Mar. 1991.
[7] X. Xu, H. Liu, Y. Li, and Y. Zhou, "Image deblurring with blur kernel estimation in RGB channels," in *Proc. IEEE Int. Conf. Digit. Signal Process. (DSP)*, Oct. 2016, pp. 681–684.
[8] T. F. Chan and C.-K. Wong, "Total variation blind deconvolution," *IEEE Trans. Image Process.*, vol. 7, no. 3, pp. 370–375, Mar. 1998.
[9] Q. Shan, J. Jia, and A. Agarwala, "High-quality motion deblurring from a single image," *ACM Trans. Graph.*, vol. 27, no. 3, p. 73, 2008.
[10] L. Xu, S. Zheng, and J. Jia, "Unnatural l0 sparse representation for natural image deblurring," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1107–1114.
[11] J. Sun, W. Cao, Z. Xu, and J. Ponce, "Learning a convolutional neural network for non-uniform motion blur removal," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 769–777.
[12] A. Chakrabarti, "A neural approach to blind motion deblurring," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 221–235.
[13] S. Anwar, Z. Hayder, and F. Porikli, "Depth estimation and blur removal from a single out-of-focus image," in *Proc. Brit. Mach. Vis. Conf.*, 2017, p. 2.
[14] 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, Jan. 2021.
[15] D. Gong *et al.*, "From motion blur to motion flow: A deep learning solution for removing heterogeneous motion blur," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2319–2328.
[16] 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. (CVPR)*, Jul. 2017, pp. 3883–3891.
[17] Q. Chen and V. Koltun, "Photographic image synthesis with cascaded refinement networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1511–1520.
[18] S. Ramakrishnan, S. Pachori, A. Gangopadhyay, and S. Raman, "Deep generative filter for motion deblurring," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 2993–3000.

[19] O. Kupyn, T. Martyniuk, J. Wu, and Z. Wang, "DeblurGAN-v2: Deblurring (orders-of-magnitude) faster and better," in *The IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8878–8887.

[20] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1125–1134.

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

[22] 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. (CVPR)*, Jun. 2021, pp. 2034–2042.

[23] S. Zhuo and T. Sim, "Defocus map estimation from a single image," *Pattern Recognit.*, vol. 44, no. 9, pp. 1852–1858, 2011.

[24] Y. Cao, S. Fang, and Z. Wang, "Digital multi-focusing from a single photograph taken with an uncalibrated conventional camera," *IEEE Trans. Image Process.*, vol. 22, no. 9, pp. 3703–3714, Sep. 2013.

[25] X. Zhang, R. Wang, X. Jiang, W. Wang, and W. Gao, "Spatially variant defocus blur map estimation and deblurring from a single image," *J. Vis. Commun. Image Represent.*, vol. 35, pp. 257–264, Feb. 2016.

[26] A. Karaali and C. R. Jung, "Edge-based defocus blur estimation with adaptive scale selection," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1126–1137, Mar. 2018.

[27] S. Liu, F. Zhou, and Q. Liao, "Defocus map estimation from a single image based on two-parameter defocus model," *IEEE Trans. Image Process.*, vol. 25, no. 12, pp. 5943–5956, Dec. 2016.

[28] J. Park, Y.-W. Tai, D. Cho, and I. S. Kweon, "A unified approach of multi-scale deep and hand-crafted features for defocus estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 487–491.

[29] P. Trouvé, F. Champagnat, G. L. Besnerais, and J. Idier, "Single image local blur identification," in *Proc. 18th IEEE Int. Conf. Image Process.*, Sep. 2011, pp. 613–616.

[30] J. Shi, L. Xu, and J. Jia, "Just noticeable defocus blur detection and estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 657–665.

[31] X. Zhu, S. Cohen, S. Schiller, and P. Milanfar, "Estimating spatially varying defocus blur from a single image," *IEEE Trans. Image Process.*, vol. 22, no. 12, pp. 4879–4891, Dec. 2013.

[32] S. Liu, Q. Liao, J.-H. Xue, and F. Zhou, "Defocus map estimation from a single image using improved likelihood feature and edge-based basis," *Pattern Recognit.*, vol. 107, Nov. 2020, Art. no. 107485.

[33] R. Yan and L. Shao, "Blind image blur estimation via deep learning," *IEEE Trans. Image Process.*, vol. 25, no. 4, pp. 1910–1921, Apr. 2016.

[34] W. Zhao, F. Zhao, D. Wang, and H. Lu, "Defocus blur detection via multi-stream bottom-top-bottom fully convolutional network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3080–3088.

[35] S. Zhang, X. Shen, Z. Lin, R. Mech, J. Costeira, and J. M. F. Moura, "Learning to understand image blur," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6586–6595.

[36] 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. (CVPR)*, Jun. 2019, pp. 12222–12230.

[37] S. Liu, A. J. Davison, and E. Johns, "Self-supervised generalisation with meta auxiliary learning," 2019, *arXiv:1901.08933*.

[38] S. Toshniwal, H. Tang, L. Lu, and K. Livescu, "Multitask learning with low-level auxiliary tasks for encoder-decoder based speech recognition," 2017, *arXiv:1704.01631*.

[39] W. Ren, J. Yang, S. Deng, D. Wipf, X. Cao, and X. Tong, "Face video deblurring using 3D facial priors," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9388–9397.

[40] L. Liebel and M. Körner, "Auxiliary tasks in multi-task learning," 2018, *arXiv:1805.06334*.

[41] J. Flynn, I. Neulander, J. Philbin, and N. Snavely, "DeepStereo: Learning to predict new views from the world's imagery," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5515–5524.

[42] A. Valada, N. Radwan, and W. Burgard, "Deep auxiliary learning for visual localization and odometry," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2018, pp. 6939–6946.

[43] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1851–1858.

[44] M. Jaderberg *et al.*, "Reinforcement learning with unsupervised auxiliary tasks," 2016, *arXiv:1611.05397*.

[45] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[46] 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 (CVPRW)*, Jul. 2017, pp. 136–144.

[47] R. Ng, M. Levoy, M. Brédif, G. Duval, M. Horowitz, and P. Hanrahan, "Light field photography with a hand-held plenoptic camera," *Comput. Sci. Tech. Rep.*, vol. 2, no. 11, pp. 1–11, 2005.

[48] D. Dansereau, O. Pizarro, and S. Williams, "Linear volumetric focus for light field cameras," *ACM Trans. Graph.*, vol. 34, no. 2, pp. 1–20, 2015.

[49] R. Ng, "Fourier slice photography," *ACM Trans. Graph.*, vol. 24, no. 3, pp. 735–744, 2005.

[50] D. G. Dansereau, B. Girod, and G. Wetzstein, "LiFF: Light field features in scale and depth," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8042–8051.

[51] A. Dosovitskiy *et al.*, "FlowNet: Learning optical flow with convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2758–2766.

[52] D. Krishnan and R. Fergus, "Fast image deconvolution using hyper-Laplacian priors," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 1033–1041.

Haoyu Ma received the B.S. and M.Eng. degrees from the Department of Electronic Engineering, Tsinghua University, China, in 2017 and 2020, respectively. His research interests include defocus imaging, image fusion, and stereo vision.

Shaojun Liu received the B.S. and Ph.D. degrees from the Department of Electronic Engineering, Tsinghua University, China, in 2014 and 2019, respectively. He is currently an Assistant Professor with the College of Health Science and Environmental Engineering, Shenzhen Technology University, China. His research interests include defocus blur identification, multi-focus image fusion, and medical image analysis.

Qingmin Liao (Senior Member, IEEE) received the B.S. degree in radio technology from the University of Electronic Science and Technology of China, China, in 1984, and the M.S. and Ph.D. degrees in signal processing and telecommunications from the University of Rennes 1, France, in 1990 and 1994, respectively. He is currently a Professor with Tsinghua Shenzhen International Graduate School, Tsinghua University. His research interests include image/video processing, transmission, and analysis, biometrics, and their applications to teledetection, medicine, industry, and sports.

Juncheng Zhang received the B.S. degree from the College of Electronics and Information Engineering, Sichuan University, China, in 2016. He is currently pursuing the Ph.D. degree in electronic engineering with Tsinghua University, China. His research interests include multi-focus image fusion and defocus blur identification.

Jing-Hao Xue (Senior Member, IEEE) received the Dr.Eng. degree in signal and information processing from Tsinghua University in 1998 and the Ph.D. degree in statistics from the University of Glasgow in 2008. He is currently a Professor with the Department of Statistical Science, University College London. His research interests include statistical classification, high-dimensional data analysis, computer vision, and pattern recognition.