NTIRE 2021 Challenge for Defocus Deblurring Using Dual-pixel Images: Methods and Results

Abdullah Abuolaim* Radu Timofte* Michael S. Brown* Dafeng Zhang Xiaobing Wang
Syed Waqas Zamir Aditya Arora Salman Khan Munawar Hayat Fahad Shahbaz Khan
Ling Shao Shuai Liu Lei Lei Chaoyu Feng Zhiwei Xiong Zeyu Xiao Ruikang Xu
Yunan Zhu Dong Liu Tu Vo Si Miao Nisarg A. Shah Pengwei Liang Zhiwei Zhong
Xingyu Hu Yiqun Chen Chenghua Li Xiaoying Bai Chi Zhang Yiheng Yao
Ruipeng Gang Sabari Nathan Thangavelu Ragavendran Venkatakrishnan Srinija
Venkatakrishnan Srivatsav

Abstract

This paper provides a review of the NTIRE 2021 challenge targeting defocus deblurring using dual-pixel (DP) data. The goal of this single-track challenge was to reduce spatially varying defocus blur present in images captured with a shallow depth of field. The images used in this challenge were obtained using a DP sensor that provided a pair of DP views per captured image. Submitted solutions were evaluated using conventional signal processing metrics, namely peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM). Out of 185 registered participants, nine teams provided methods and competed in the final stage. The paper describes the methods proposed by the participating teams and their results. The winning teams represent the state-of-the-art in terms of defocus deblurring using DP images.

1. Introduction

Defocus blur occurs in an image at scene points when the light rays, traveling through the camera optics, converge either before or after the imaging sensor. While the effect of defocus blur can be intentional in photography (*e.g.*, the bokeh effect [13, 32]), for computer vision applications, defocus blur is often undesired and affects image quality due to the loss of sharp image details. Recovering sharper details from a defocus deblurred image is challenging [15] due to the spatially varying nature of blur shape and size because of optical aberrations across scene depth [30].

Recently, Abuolaim *et al*. [1] demonstrated the advantage of utilizing dual-pixel (DP) data to reduce defocus blur. The DP sensor was designed by Canon to assist phase-difference autofocus mechanism used in many modern cameras. In particular, the DP sensor is constructed with two photodiodes at each pixel location. When capturing an image, the two photodiodes allow the capture of two sub-aperture views of the same scene in a single capture. The DP views have a difference in phase that is correlated to the amount of defocus blur. When a DP sensor is coupled with an adjustable lens, an autofocus algorithm can quickly correct the lens position to minimize the phase difference between the DP views and bring scene content into focus. While intended for camera autofocus [2], DP data was found to be useful for other computer vision applications *e.g.*, depth map estimation [10, 26, 41], synthetic bokeh [32], and defocus deblurring [1].

This challenge is one of the NTIRE 2021 associated challenges: nonhomogeneous dehazing [4], defocus deblurring using dual-pixel [3], depth guided image relighting [8], image deblurring [22], multi-modal aerial view imagery classification [17], learning the super-resolution space [20], quality enhancement of heavily compressed videos [37], video super-resolution [29], perceptual image quality assessment [11], burst super-resolution [5], high dynamic range [24]. This single-track challenge solicited algorithms that can reduce defocus blur using DP data. The challenge, dataset, submitted solutions, and the results based on PSNR and SSIM are described in the subsequent sections.

*A. Abuolaim (abuolaim@eecs.yorku.ca, York University), R. Timofte (radu.timofte@vision.ee.ethz.ch, ETH Zurich), and M.S. Brown (mbrown@eecs.yorku.ca, York University) are the organizers of the NTIRE 2021 defocus deblurring challenge. The rest of the author list are the participants of the teams who competed in the final stage.
Appendix A provides the team names and affiliations of the participants.
NTIRE 2021 webpage:
https://data.vision.ee.ethz.ch/cvl/ntire21/

2. The Challenge

The NTIRE 2021 challenge on defocus deblurring using DP images is aimed to gauge and advance the state-of-the-art in reducing defocus blur. This challenge aims to evaluate the performance of the proposed defocus-deblurring methods on a newly captured DP dataset. The following provides a detailed description of the DP datasets used, the evaluation procedure and metrics, and the challenge timeline.

2.1. Datasets

The dataset used in this challenge consists of images carefully captured on a tripod with the aperture adjusted between image captures. Two images are captured for the same static scene in succession; the first is a wide-aperture image captured using f/4 and exhibits notable defocus blur, while the second is captured with a narrow aperture (*i.e.*, f/22) and exhibits a wide depth of field with almost no defocus blur. The narrow aperture image serves as the ground truth sharp image.

The training data used for this challenge consists of the 500 indoor/outdoor scenes from the DP defocus deblurring dataset[1] [1]. In addition to this dataset, we captured a new 100 indoor/outdoor scenes divided equally for the validation and testing data. The challenge provides 600 scenes in total (*i.e.*, 1800 images) where each scene has: (i) the two DP sub-aperture views of an image with defocus blur captured at a large aperture; and (ii) the corresponding all-in-focus image captured with a small aperture.

The dataset images are high-quality as they are captured with low ISO (*i.e.*, low ISO equates to low-noise [25]) and have a $1,680 \times 1,120$ spatial resolution. These images, including the left/right DP views, are processed to an sRGB color space and encoded with a lossless 12-bit depth per RGB channel.

2.2. Evaluation

The evaluation compares the recovered sharp (*i.e.*, deblurred) images with the ground-truth images. For this comparison, we use the standard peak signal-to-noise ratio (PSNR) and the structural similarity (SSIM) index [34] as often employed in the literature. We report the average results over all the estimated test images provided.

Challenge participants were asked to provide the estimated deblurred images with the same resolution as the input and using the specified naming convention. Participants were also asked to provide additional information, for example, the algorithm's runtime per test image (in seconds); whether the algorithm employs CPU or GPU at runtime, and whether extra metadata is used as inputs to the algorithm.

[1] https://www.eecs.yorku.ca/~abuolaim/eccv_2020_dp_defocus_deblurring/

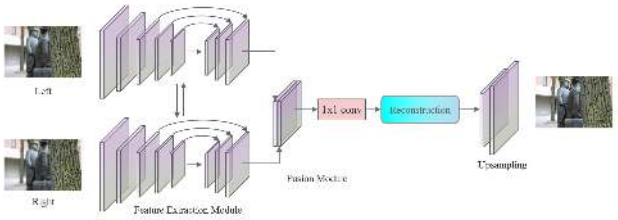

Figure 1. Network architecture of multi-refinement network for dual-pixel images defocus deblurring (MRNet).

At the final stage of the challenge, the participants were asked to submit fact sheets to provide information about the teams and describe their methods. In addition to the fact sheets, the output results, codes and trained models were also submitted.

2.3. Timeline

The challenge timeline had two stages – validation and testing. The validation stage commenced on January 6, 2021, and lasted for approximately 10 weeks. The final testing stage started on March 15, 2021, and lasted for 5 days. Each participant was allowed 20 submissions during the validation stage and three submissions for the testing phase. The challenge ended on March 20, 2021. The final test results were shared with the participants on March 26, 2021.

3. Proposed Methods

This section reviews the details of the proposed methods, where the description of each method is provided by the team members.

3.1. SRC-B Team

The SRC-B team proposed the MRNet: Multi-Refinement Network for Dual-pixel Images Defocus Deblurring, as shown in Fig. 1. The proposed DP images defocus deblurring neural network is mainly composed of 4 modules: feature extraction module, fusion module, reconstruction module and upsampling module.

The feature extraction module mainly uses Siamese network [14] to extract features of left and right input images, in which the Siamese network is the weight sharing. The original letf and right images are downsampled 16 times for aligning in the spatial position that cause by the position deviation of the left and right sensors. Downsampling will lose a lot of detailed information, so then upsample the features by 4 times. On the one hand, it protects more detailed information of the input features, and on the other hand, it can speed up the network.

We use a simple way to fuse the left and right features extracted by the feature extraction module. We use the CONCAT operation to concatenate left and right features in the

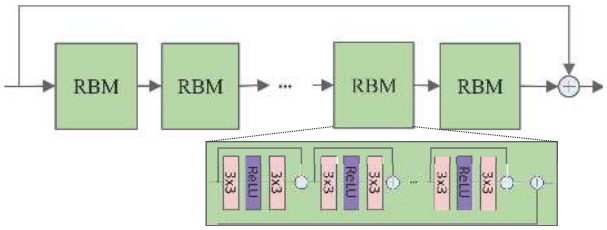

Figure 2. Residual Group Module. RBM: is the a single residual block module.

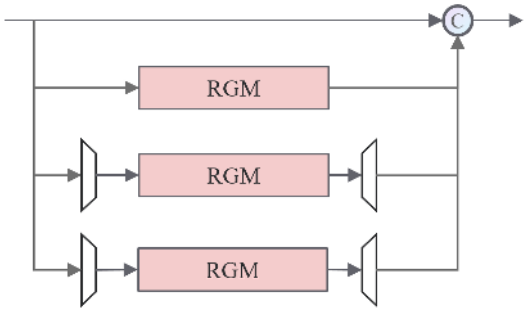

Figure 3. Multi-scale Residual Group Module.

Figure 4. Reconstruction Module.

channel dimension. Then, the convolved kernel with size 1 is used to reduce its dimension to get the fused features.

MMDM [19] verified the effectiveness of the proposed feature extraction and reconstruction module (FERM) in the NTIRE 2020 Challenge on the Image Demoireing track. Inspired by MMDM, we proposed Residual Block Module (RBM) as the basic module of the reconstructed module. RBM adopts the same configuration as MMDM, which consists of 10 residual modules and a global residual connection. And like MMDM, we don't use Channel Attention (CA) module. CA module will increase the inference and training time of the model, but the benefit is very small. MMDM uses 20 RBM modules and many 3x3 convolution modules to form FERM module, while we only use 5 RBM modules to form RGM module to accelerate the training and inference speed of the network. RBM and RGM are shown in Fig. 2. To process the various frequency components in the moire patterns, MMDM proposes a multi-scale feature encoding module (MSFE) that processes images at different scales. The MSFE has 3 simple versions of FERM with up and downsampling layers for different scales. We also proposed the Multi-scale Residual Group Module (MSRGM) to fuse features of different scales to improve the model's expressive ability. But in the training phase, we use a patch

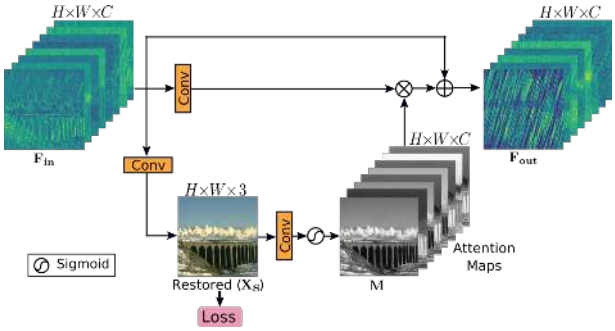

Figure 5. Supervised attention module (SAM).

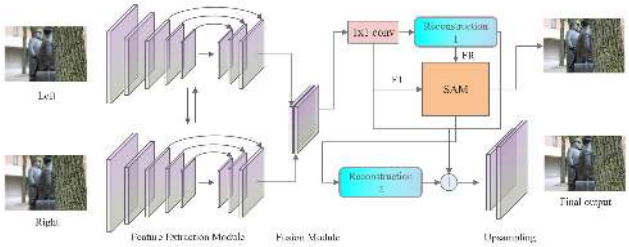

Figure 6. MRNet-SAM.

size of 512x512, and the feature dimension entered in the reconstruction module is 128x128. 8 times downsampling operation will result in very small feature size, which will lose a lot of detail information. Therefore, we use the module shown in the Fig. 3. to extract multi-scale features. The reconstruction module (RM) is composed of multiple MSRGM modules, as shown in the Fig. 4. The RM, RGM, and RBM modules all add global residual connections. The idea is that each block is a refinement of the previous feature. Especially in RM, we fuse the multi-scale features and refine them in the next MSRGM.

Refinement is the core idea of their method. In MRNET, we use MSRGM to fuse and refine the features from different scales. Inspired by MPRNet(CVPR2021) [39], we improve MRNet and propose a cross-stage progressive neural network based on supervised attention module (SAM), which we named MRNet-SAM. MRNet-SAM is shown in Fig. 6. MPRNet introduced a supervised attention module (SAM) between every two stages, facilitating significant performance gain. We modified SAM to make it more suitable for the task due to the different sizes between input and output features. SAM is shown in Fig. 5. We divide the reconstruction module in MRNet into two parts as the reconstruction module 1 and reconstruction module 2 in MRNet-SAM, and add the SAM module to obtain better performance. At the same time, we replace the convolution with stride 2 in the MSRGM with the DownPixelShuffle operation. Compared with convolution with stride 2 methods, DownPixelShuffle can preserve the information and reduce

the parameters.

3.2. TeamInception Team

The TeamInception team presents a Multi-Stage Progressive Image Restoration MPRNet that is recently introduced in [39]. As illustrated in Fig. 7, MPRNet consists of three stages to progressively restore images. The first two stages are based on encoder-decoder subnetworks that learn the broad contextual information due to large receptive fields. Since defocus deblurring is a position-sensitive task (which requires pixel-to-pixel correspondence from the input to output), the last stage employs a subnetwork, named ORRNets, that operates on the original input image resolution (without any downsampling operation), thereby preserving the desired fine texture in the final output image. ORSNet contains multiple original resolution blocks (ORBs). ORB is shown in Fig. 8.

Instead of simply cascading multiple stages, we incorporate a supervised attention module (SAM) between every two stages. The schematic diagram of SAM is shown in Fig. 5. Our model rescales the feature maps of the previous stage with the supervision of ground-truth images before passing them to the next stage. Furthermore, we introduce a cross-stage feature fusion mechanism where the intermediate multi-scale contextualized features of the earlier subnetwork help consolidate the latter subnetwork's intermediate features.

Although MPRNet stacks multiple stages, each stage has access to the input image. We adapt the multi-patch hierarchy on the input image and split the image into non-overlapping patches: four for stage-1, two for stage-2, and the original image for the last stage, as shown in Fig. 7. Furthermore, the restored image at each stage is concatenated to the next stage. For more architectural details, we refer the interested readers to [39].

Loss Function. To optimized the proposed network, we use the following loss function.

$$\mathcal{L}_f = \alpha \mathcal{L}_1(\hat{\mathbf{y}}, \mathbf{y}) + \beta \mathcal{L}_{\text{MS-SSIM}}(\hat{\mathbf{y}}, \mathbf{y}) + \gamma \mathcal{L}_{\text{VGG}}(\hat{\mathbf{y}}, \mathbf{y}) \quad (1)$$

The first term (L1 loss) and second term (multi-scale structural similarity measure) compute differences between the network's output and the ground truth directly at the pixel level. The last term of the loss function compares the deep feature representations of the output and ground-truth images extracted with the VGG network pre-trained on the ImageNet dataset [28]. In Fig. 7 we show the framework of our MPRNet.

$$\mathcal{L}_{VGG}(\hat{\mathbf{y}}, \mathbf{y}) = \frac{1}{N} \parallel \phi(\hat{\mathbf{y}}) - \phi(\mathbf{y}) \parallel_2^2, \quad (2)$$

where N denotes the total number of pixels in the image. In our experiments we use *conv2* layer after ReLU of the VGG-16 network.

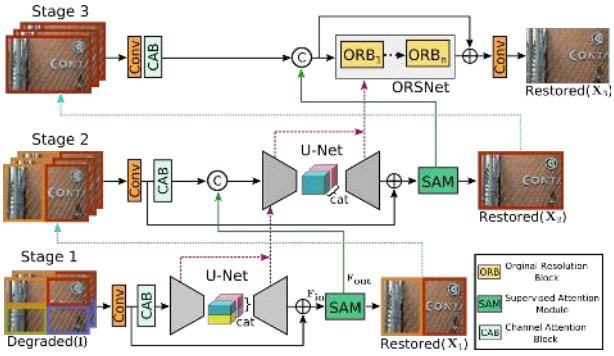

Figure 7. TeamInception team. The proposed multi-stage architecture for progressive image restoration (MPRNet). Earlier stages employ encoder-decoders to extract multi-scale contextualized features, while the last stage operates at the original image resolution to generate spatially accurate outputs. A supervised attention module (SAM) is added between every two stages that learns to refine features of one stage before passing them to the next stage. Dotted pink arrows represent the cross-stage feature fusion mechanism.

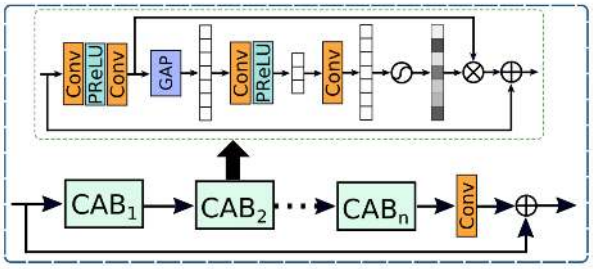

Figure 8. Illustration of the original resolution block (ORB) in our ORSNet subnetwork. Each ORB contains multiple channel attention blocks (CABs). GAP represents global average pooling.

3.3. Mier Team

The Mier team proposed a Big UNet for image restoration based on both MWCNN [18] and RCAN [40]. They replaced the convolutional layer in MWCNN with the residual group (RG, without channel attention layer) in RCAN to enhance the reconstruction ability of the network. In order to further expand the receptive field, they also added a multi-scale dilate block (MDB [42]) to the network. The network structure is shown in the Fig. 9.

3.4. VIDAR Team

The VIDAR team propose a multi-scale parallax attention network with a two-stage structure as shown in Fig. 10. In stage 1, we adopt a multi-scale feature extractor to extract discriminative features for the following processes. Then, we design a multi-scale parallax attention module to fully exploit the correlation of DP images at different scales. Specifically, inspired by [33], we modify the

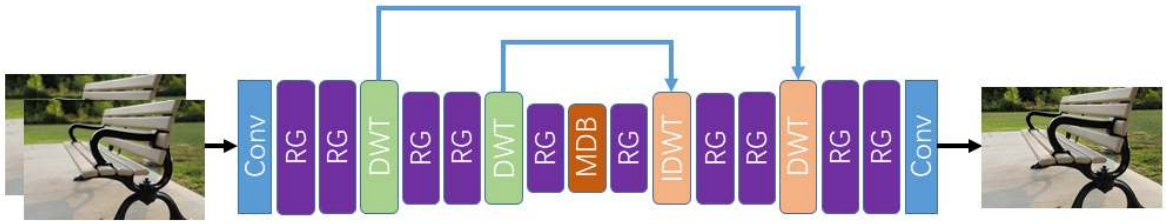

Figure 9. The Big UNet architecture proposed by the Mier team

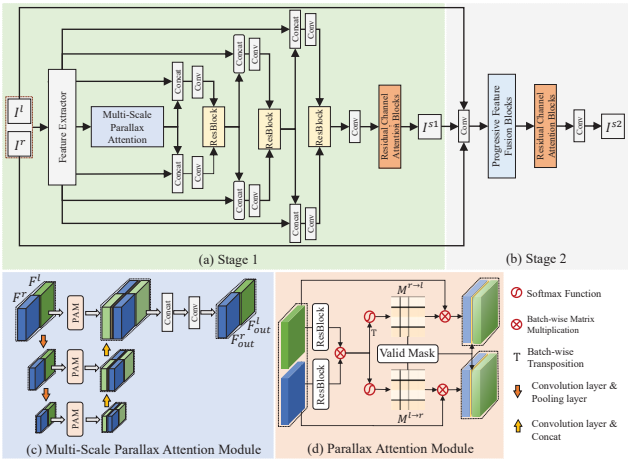

Figure 10. The multi-scale parallax attention network proposed by the VIDAR team. I^l and I_r are the defocus blurred left and right images. I^{s1} and I^{s2} are the results of stage 1 and stage 2, respectively. F^l, F^r, F^l_{out} and F^r_{out} are the input left feature, the input right feature, the output left feature and the output right feature of the multi-scale parallax attention module. $M^{r\to l}$ and $M^{l\to r}$ are parallax attention maps generated by the parallax attention module.

parallax-attention mechanism with a pyramid structure to fully exploit the parallax information and stereo correspondence from DP images with enlarged receptive fields. This module can generate stereo-symmetric features and aggregate the information from DP data. Finally, we use 20 residual channel attention blocks [40] as the reconstruction backbone. In stage 2, the output of stage 1 and the original input DP images are fed into a convolution layer and 10 progressive feature fusion blocks [38]. Then, the fused features are fed into 20 residual channel attention blocks to generate the high-quality image.

3.5. Attention Team

The Attention team proposed an algorithm to exploit the properties of each pixel and each channel of the input images to generate the non-blurred output image as shown in Fig. 11. We designed a network architecture that is an Unet-liked with dual-attention modules, which are employed in every down-scaling level. Furthermore, the model also contains the global non-local modules, which ensure to make the network intentionally learns the local patterns without losing the global image structures, which makes the output images to be closer to the real human visual system.

Attention Encoder. The attention encoder is shown in Fig. 12. Our encoder takes the advantages of the dual-attention modules [9] where the *sigmoid* function is employed in both channel-wise and pixel-wise, which helps the encoder to learn the useful information from both input images. Our attention encoder modules take the output of the previous module; the dual-attention module will selectively select the channel as well as the pixel to encode and move to the next module. The *sigmoid* function ensures that the useful information will be passed through while the others will be discarded.

Triplet Local. Since it is beneficial to learn the feature in different local levels, we designed the triplet local modules, which uses different kernel sizes. Different kinds of details can be learned from different receptive fields. The small kernel size devoted to extracting local features and the larger kernel size can cover more extensive regions of the receiving layers.

Global Non-Local We calculate the non-local information in the feature domain by firstly using adaptive pooling to 7×7 features in the global non-local module; then the global non-local will selectively select the region to be fused. Note that the information of the output of this module covers the information of the entire input. The output is then adaptively upsampled to the output size of the triplet local.

3.6. Maradona Team

Maradona Team follow Nah's contribution [21]. It is a two-scale network (see Fig. 13). We use 28 Resblocks[12] in the low-resolution branch and 35 Dilation blocks [6, 42] in the high-resolution branch. The low-resolution result is up-sampled to the higher resolution that concats the original blurred dual images on the channel.

3.7. AIIA Team

The AIIA team proposed a Multi-Branch Convolution Neural Network for Dual-Pixel Image Deblurring. The main network architecture is a set of autoencoders with an

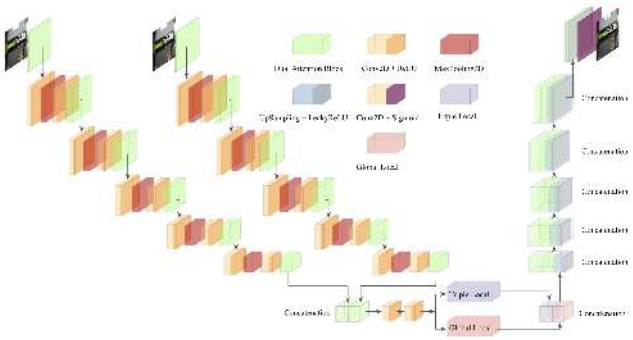

Figure 11. Attention team proposed architecture. We redesign the original encoder by adding the Dual-Attention Module on top of it. This ensures the encoders extract useful information wisely at every pixel location as well as channel position.

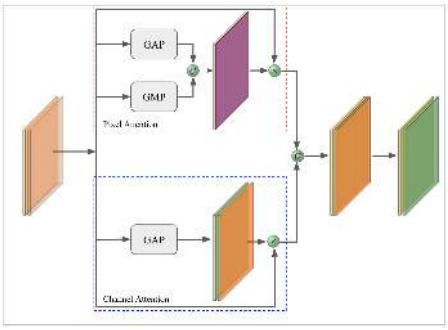

Figure 12. Attention team. Dual Attention Module consists of two parallel sub-modules, which help the encoder decide the level of contribution for each pixel and each channel of the input feature.

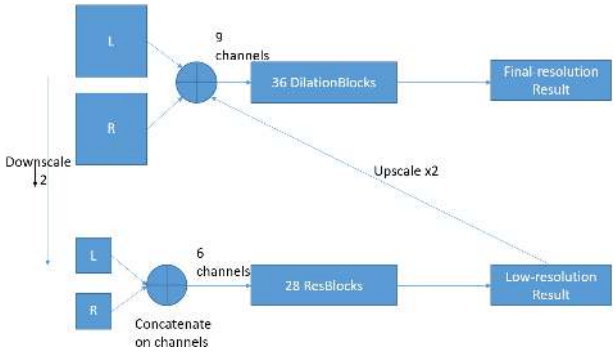

Figure 13. An overview of Maradona team proposed method.

assignment module. To train the model, a patch filter is used to convert 1680×1120 size images to 512×512 size patches. The model is optimized using the Adam optimizer to minimize the MSE loss. The initial learning rate is 2×10^5. AIIA team proposed a novel multi-branch network with an assignment module to assign different branch networks to different pixels for DP image deblurring.

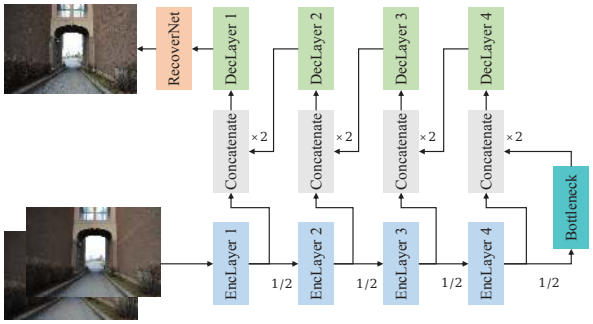

Figure 14. Overview of our model. Each encoder layer (EncLayer i, where $i \in \{1, 2, 3, 4\}$) and decoder layer (DecLayer i) consists of two convolution layers. $1/2$ and $\times 2$ represent maxpooling which halve the feature size and nearest upsampling which double feature size, respectively. Both "Bottleneck" and "RecoverNet" are simple networks which contain two convolution layers.

3.8. DDDP Team

Inspired by [1], the proposed method takes an encoder-decoder fashion, which consists of four parts: 1) Encoder; 2) Decoder; 3) Bottleneck; 4) RecoverNet, as shown in Figure 14. We also develop a novel data augmentation method to impose a model to learn spatial relations.

We also treat the ground truth as left patches and right patches in addition to the original data. By introducing this data augmentation technique, we can easily double the data scale and interpret a model as a spatial relation learner.

We adopt a loss function which is the weighted sum of MSE loss and SSIM loss. By denoting the output of the proposed method and the ground-truth as O and Y, respectively, the total loss can be formulated as:

$$Loss = \lambda_1 \cdot \text{MSE}(O, Y) + \lambda_2(1 - \text{SSIM}(O, Y)) \quad (3)$$

where λ_1 and λ_2 are the weight of MSE loss and SSIM loss, respectively.

3.9. ImageLab Team

The ImageLab team proposed network has two parallel encoders and a decoder as shown in Fig 15. The Input Image I_a and Input Image I_b are being passed into each DualVisionNet encoder. Encoder block has the convolution layer with a 3x3 filter followed by two Densely Connected Recurrent Residual block [7], and the end of the block max-pooling layer is added. Two encoder outputs are concatenated and passed to the convolutional block attention module (CBAM) [36] attention layer. The decoder uses the same block as an encoder, but the max-pooling layers are replaced with the upsampling layers. Output features of encoder $block_x$, encoder $block_y$ are concatenated with the $decoder_X$ upsampled features . In this network, the encoder acts as a fusion network that helps us find the pair of

Table 1. Results and rankings by team of the submitted methods for NTIRE 2021 defocus deblurring challenge. MS-SSIM [35]: multi-scale structural similarity loss. The runtime of s/Mpixel is calculated based on the runtime of a single test image of size $1680 \times 1120 \times 3$ pixels (*i.e.*, 5.6448 Mpixel).

Team	Username	PSNR	SSIM	Runtime (s/Mpixel)	CPU/GPU (at runtime)	Platform	Ensemble	Loss
SRCB	SRCB-Z	$27.80_{(1)}$	$0.8484_{(1)}$	$0.029_{(1)}$	Tesla V100	PyTorch	flip, multi-model ensemble	Charbonnier
TeamInception	swz30	$27.48_{(2)}$	$0.8387_{(3)}$	$0.514_{(7)}$	Tesla V100	PyTorch	flip, rotate, self-ensemble [31]	L_1, VGG [28], MS-SSIM [35]
Mier	q935970314	$27.13_{(3)}$	$0.8408_{(2)}$	$10.62_{(9)}$	Tesla V100	PyTorch	left/right swap, self-ensemble [31]	L_1
VIDAR	zyxiao	$26.79_{(4)}$	$0.8264_{(4)}$	$0.046_{(3)}$	Tesla V100	PyTorch	flip, inverse transform	Charbonnier, SSIM
Attention	buffalo	$26.42_{(5)}$	$0.8020_{(6)}$	$0.097_{(5)}$	Tesla V100	Tensorflow/Keras	flip/rotate	L_2, SSIM
Maradona	hellosr	$26.38_{(6)}$	$0.8017_{(7)}$	$08.70_{(8)}$	Tesla V100	PyTorch	Geometric self-ensemble$\times 8$ [16]	L_1
AIIA	huxingyu	$26.15_{(7)}$	$0.8021_{(5)}$	$0.482_{(6)}$	RTX 3090	PyTorch	None	L_2
DDDP	BaiXiaoying	$25.17_{(8)}$	$0.7620_{(8)}$	$0.081_{(4)}$	TITAN Xp	PyTorch	None	L_2, SSIM
ImageLab	sabarinathan	$24.85_{(9)}$	$0.7390_{(9)}$	$0.041_{(2)}$	1080 GTX	Tensorflow/Keras	None	L_2, SSIM, Sobel

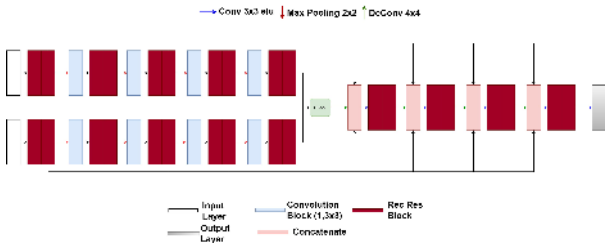

Figure 15. ImageLab team proposed model. A multi-level attention based efficient U-Net encoder decoder model

high frequency and low-frequency components from the I_a and I_b. The Loss function is inspired by Multi-Level Hyper Vision Net [23] model.

$$Loss = MSE + (1 - SSIM) + SOBEL_{loss}$$

4. Results

Out of 185 registered participants, the challenge had nine teams who continued to the final stage by submitting results, codes/executables, and fact sheets. Tables 1 reports the final test results based on PSNR and SSIM index [34]. The overall method rank based on each metric is indicated in subscripts. Additionally, the self-reported runtimes and major details, provided in the fact sheets submitted by participants, are also reported in Tables 1. Fig 16 show a 2D visualization of PSNR and SSIM values for all teams. The team members along with affiliations are listed in Appendix A.

4.1. Core Idea

All of the proposed methods are based on deep learning. Particularly, all methods employ different architectures of deep convolutional neural networks (CNN). Most of proposed architectures are based on widely used CNN-based networks, such as Siamese network [14], U-Net [27] and ResNet [12]. The core ideas included re-structuring existing networks, introducing skip connections, introducing residual connections, and using densely connected components. Other strategies have been utilized including supervised attention module (SAM) [39], residual channel attention blocks [40], dual-attention modules [9], convolutional block attention module (CBAM) [36], and dilation blocks [6, 42].

As for loss functions, different types were employed such as L_1, L_2, Charbonnier, SSIM, multi-scale SSIM [35] (MSSSIM), VGG [28], and Sobel. Some teams used a single loss, whereas others used a mix of loss functions (*e.g.*, TeamInception utilized L_1, VGG [28], and MSSSIM [35]).

4.2. Top Results

The SRCB team has achieved the best results for all metrics, including the inference time (*i.e.*, 0.029 s/Mpixel).

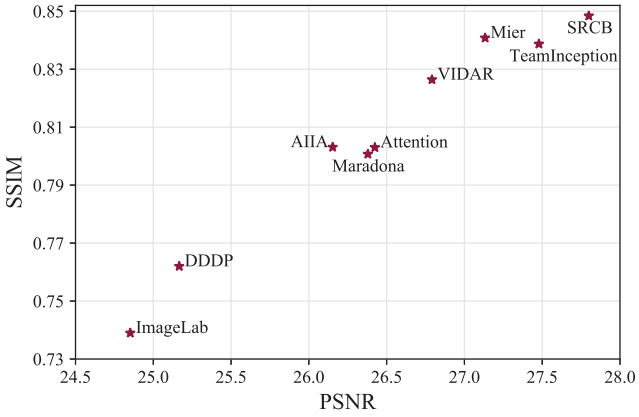

Figure 16. A 2D visualization of the combined PSNR and SSIM values of the proposed methods. Each team name is shown next to the scattered star markers.

They are 0.32 dB higher compared to the second top team *i.e.*, TeamInception as reported in Table 1. In terms of PSNR, the main performance metric used in the challenge, the top three methods achieved larger than 27dB, and they are from the teams of SRCB, TeamInception, and Mier, respectively (see Fig. 16). In terms of SSIM, as a complementary performance metric, the second-best method is proposed by the Mier team and achieved a SSIM index of 0.8408, while the third-best SSIM index is TeamInception team, *i.e.*, 0.8387.

4.3. Ensembles

Most of the teams applied different flavors of ensemble techniques to boost the overall performance. In particular, most teams used a self-ensemble [31] technique where the results from flipped/rotated versions of the same image are averaged together. Some teams applied additional techniques such as inverse transform, geometric self-ensemble [16], and multi-model ensemble.

5. Conclusion

In this paper, we reviewed the methods submitted for the NTIRE 2021 challenge for defocus deblurring using DP images with focus on the methods and their results. In particular, an evaluation of methods proposed by nine teams is performed based on PSNR, SSIM, and inference time. In addition to the DP deblurring dataset from [1], we provided a new DP-based deblurring dataset for further evaluations. All the proposed methods are based on deep CNNs and most of them utilized attention layers to boost performance. The best performing method is proposed by the SRCB team from Samsung research Beijing, China, where they achieved the best results for all metrics.

Acknowledgements

We thank the NTIRE 2021 sponsors: Huawei, Facebook Reality Labs, Wright Brothers Institute, MediaTek, and ETH Zurich (Computer Vision Lab).

A. Team Names and Affiliations

NTIRE 2021 Organization

Title: NTIRE 2021 Challenge for Defocus Deblurring Using Dual-pixel Images: Methods and Results
Members:
Abdullah Abuolaim[1] (abuolaim@eecs.yorku.ca),
Radu Timofte[2] (radu.timofte@vision.ee.ethz.ch),
Michael S. Brown[1] (mbrown@eecs.yorku.ca)
Affiliations:
[1] York University, Canada
[2] ETH Zurich, Switzerland

Samsung Research China – Beijing (SRC-B)

Title: MRNet: Multi-Refinement Network for Dual-pixel Images Defocus Deblurring
Members:
Dafeng Zhang (dfeng.zhang@samsung.com), Xiaobing Wang
Affiliations:
Samsung Research China – Beijing (SRC-B), China

TeamInception

Title: Multi-Stage Progressive Image Restoration
Members:
Syed Waqas Zamir[1] (waqas.zamir@inceptioniai.org), Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ling Shao
Affiliations:
[1] Inception Institute of Artificial Intelligence (IIAI), UAE

Mier

Title: Big UNet
Members:
Shuai Liu[1] (18601200232@163.com), Lei Lei[2] , Chaoyu Feng[2]
Affiliations:
[1] North China University of Technology, China
[2] Xiaomi, China

VIDAR

Title: Multi-Scale Feature Fusion Network for Defocus Deblurring Using Dual-Pixel Images
Members:
Zhiwei Xiong (zwxiong@ustc.edu.cn), Zeyu Xiao, Ruikang Xu, Yunan Zhu, Dong Liu
Affiliations:
University of Science and Technology of China, China

Attention

Title: Attention! Stay Focus!
Members:
Tu Vo (tuvovan@pukyong.ac.kr)
Affiliations:
Bridge AI Inc., Korea

Maradona

Title: Multi-scale CNN for Dual Defocus Deblurring
Members:
Si Miao[1] (miaosi2018@sari.ac.cn), Nisarg A. Shah[2]
Affiliations:
[1]Shanghai Advanced Research Institute, Chinese Academy of Sciences, China
[2]Indian Institute of Technology Jodhpur, India

AIIA

Title: Multi-Branch Convolution Neural Network for Dual-Pixel Image Deblurring
Members:
Pengwei Liang[1] (erfect@stu.hit.edu.cn), Zhiwei Zhong, Xingyu Hu
Affiliations:
[1]Harbin Institute of Technology, China

DDDP

Title: Defocus Deblurring via Dual-pixel Images
Members:
Yiqun Chen[1] (chenyiqun2021@ia.ac.cn), Chenghua Li, Xiaoying Bai, Chi Zhang, Yiheng Yao, Ruipeng Gang
Affiliations:
[1] Institute of Automation, Chinese Academy of Sciences, China

ImageLab

Title: A Dual Vision Net : A novel approach for Deblurring the Dual-pixel Images
Members:
Sabari Nathan[1] (sabarinathantce@gmail.com), Thangavelu Ragavendran[2], Venkatakrishnan Srinija[3], Venkatakrishnan Srivatsav[4]
Affiliations:
[1]Couger Inc., India
[2]Jeppair College of Engineering, India
[3]Sri Sai Ram Institute of Technology, India
[4]Prince Shri Venkateshwara Padmavathy Engineering College, India

References

[1] Abdullah Abuolaim and Michael S Brown. Defocus deblurring using dual-pixel data. In *ECCV*, 2020. 1, 2, 6, 8

[2] Abdullah Abuolaim, Abhijith Punnappurath, and Michael S Brown. Revisiting autofocus for smartphone cameras. In *ECCV*, 2018. 1

[3] Abdullah Abuolaim, Radu Timofte, Michael S Brown, et al. NTIRE 2021 challenge for defocus deblurring using dual-pixel images: Methods and results. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 1

[4] Codruta O Ancuti, Cosmin Ancuti, Florin-Alexandru Vasluianu, Radu Timofte, et al. NTIRE 2021 nonhomogeneous dehazing challenge report. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 1

[5] Goutam Bhat, Martin Danelljan, Radu Timofte, et al. NTIRE 2021 challenge on burst super-resolution: Methods and results. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 1

[6] Stephan Brehm, Sebastian Scherer, and Rainer Lienhart. High-resolution dual-stage multi-level feature aggregation for single image and video deblurring. In *CVPR workshops*, 2020. 5, 7

[7] Kaushik Dutta. Densely connected recurrent residual (dense r2unet) convolutional neural network for segmentation of lung ct images. *arXiv preprint arXiv:2102.00663*, 2021. 6

[8] Majed El Helou, Ruofan Zhou, Sabine Süsstrunk, Radu Timofte, et al. NTIRE 2021 depth guided image relighting challenge. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 1

[9] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *CVPR*, 2019. 5, 7

[10] Rahul Garg, Neal Wadhwa, Sameer Ansari, and Jonathan T Barron. Learning single camera depth estimation using dual-pixels. In *ICCV*, 2019. 1

[11] Jinjin Gu, Haoming Cai, Chao Dong, Jimmy S. Ren, Yu Qiao, Shuhang Gu, Radu Timofte, et al. NTIRE 2021 challenge on perceptual image quality assessment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 1

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5, 7

[13] Andrey Ignatov, Jagruti Patel, and Radu Timofte. Rendering natural camera bokeh effect with deep learning. In *CVPR workshops*, 2020. 1

[14] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2, 2015. 2, 7

[15] Anat Levin, Yair Weiss, Fredo Durand, and William T Freeman. Understanding blind deconvolution algorithms. *TPAMI*, 33(12):2354–2367, 2011. 1

[16] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *CVPR workshops*, 2017. 7, 8

[17] Jerrick Liu, Oliver Nina, Radu Timofte, et al. NTIRE 2021 multi-modal aerial view object classification challenge. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 1

[18] Pengju Liu, Hongzhi Zhang, Kai Zhang, Liang Lin, and Wangmeng Zuo. Multi-level wavelet-cnn for image restoration. In *CVPR workshops*, 2018. 4

[19] Shuai Liu, Chenghua Li, Nan Nan, Ziyao Zong, and Ruixia Song. Mmdm: Multi-frame and multi-scale for image demoiréing. In *CVPR workshops*, 2020. 3

[20] Andreas Lugmayr, Martin Danelljan, Radu Timofte, et al. NTIRE 2021 learning the super-resolution space challenge. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 1

[21] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *CVPR*, 2017. 5

[22] Seungjun Nah, Sanghyun Son, Suyoung Lee, Radu Timofte, Kyoung Mu Lee, et al. NTIRE 2021 challenge on image deblurring. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 1

[23] D Sabari Nathan, M Parisa Beham, and SM Roomi. Moire image restoration using multi level hyper vision net. *arXiv preprint arXiv:2004.08541*, 2020. 7

[24] Eduardo Pérez-Pellitero, Sibi Catley-Chandar, Aleš Leonardis, Radu Timofte, et al. NTIRE 2021 challenge on high dynamic range imaging: Dataset, methods and results. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 1

[25] Tobias Plotz and Stefan Roth. Benchmarking denoising algorithms with real photographs. In *CVPR*, 2017. 2

[26] Abhijith Punnappurath, Abdullah Abuolaim, Mahmoud Afifi, and Michael S Brown. Modeling defocus-disparity in dual-pixel sensors. In *ICCP*, 2020. 1

[27] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 7

[28] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 4, 7

[29] Sanghyun Son, Suyoung Lee, Seungjun Nah, Radu Timofte, Kyoung Mu Lee, et al. NTIRE 2021 challenge on video super-resolution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 1

[30] Huixuan Tang and Kiriakos N Kutulakos. Utilizing optical aberrations for extended-depth-of-field panoramas. In *ACCV*, 2012. 1

[31] Radu Timofte, Rasmus Rothe, and Luc Van Gool. Seven ways to improve example-based single image super resolution. In *CVPR*, 2016. 7, 8

[32] Neal Wadhwa, Rahul Garg, David E Jacobs, Bryan E Feldman, Nori Kanazawa, Robert Carroll, Yair Movshovitz-Attias, Jonathan T Barron, Yael Pritch, and Marc Levoy. Synthetic depth-of-field with a single-camera mobile phone. *ACM Transactions on Graphics*, 37(4):64, 2018. 1

[33] Longguang Wang, Yingqian Wang, Zhengfa Liang, Zaiping Lin, Jungang Yang, Wei An, and Yulan Guo. Learning parallax attention for stereo image super-resolution. In *CVPR*, 2019. 4

[34] Zhou Wang, Alan C Bovik, Hamid R Sheikh, Eero P Simoncelli, et al. Image quality assessment: from error visibility to structural similarity. *TIP*, 13(4):600–612, 2004. 2, 7

[35] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pages 1398–1402. IEEE, 2003. 7

[36] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *ECCV*, 2018. 6, 7

[37] Ren Yang, Radu Timofte, et al. NTIRE 2021 challenge on quality enhancement of compressed video: Methods and results. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 1

[38] Peng Yi, Zhongyuan Wang, Kui Jiang, Junjun Jiang, and Jiayi Ma. Progressive fusion video super-resolution network via exploiting non-local spatio-temporal correlations. In *ICCV*, 2019. 5

[39] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. In *CVPR*, 2021. 3, 4, 7

[40] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *ECCV*, 2018. 4, 5, 7

[41] Yinda Zhang, Neal Wadhwa, Sergio Orts-Escolano, Christian Häne, Sean Fanello, and Rahul Garg. Du2net: Learning depth estimation from dual-cameras and dual-pixels. *ECCV*, 2020. 1

[42] Shangchen Zhou, Jiawei Zhang, Wangmeng Zuo, Haozhe Xie, Jinshan Pan, and Jimmy S Ren. Davanet: Stereo deblurring with view aggregation. In *CVPR*, 2019. 4, 5, 7