Learning Enriched Features for Fast Image Restoration and Enhancement

Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, *Fellow, IEEE*, and Ling Shao, *Fellow, IEEE*

Abstract—Given a degraded input image, image restoration aims to recover the missing high-quality image content. Numerous applications demand effective image restoration, e.g., computational photography, surveillance, autonomous vehicles, and remote sensing. Significant advances in image restoration have been made in recent years, dominated by convolutional neural networks (CNNs). The widely-used CNN-based methods typically operate either on full-resolution or on progressively low-resolution representations. In the former case, spatial details are preserved but the contextual information cannot be precisely encoded. In the latter case, generated outputs are semantically reliable but spatially less accurate. This paper presents a new architecture with a holistic goal of maintaining spatially-precise high-resolution representations through the entire network, and receiving complementary contextual information from the low-resolution representations. The core of our approach is a multi-scale residual block containing the following key elements: (a) parallel multi-resolution convolution streams for extracting multi-scale features, (b) information exchange across the multi-resolution streams, (c) non-local attention mechanism for capturing contextual information, and (d) attention based multi-scale feature aggregation. Our approach learns an enriched set of features that combines contextual information from multiple scales, while simultaneously preserving the high-resolution spatial details. Extensive experiments on six real image benchmark datasets demonstrate that our method, named as MIRNet-v2 , achieves state-of-the-art results for a variety of image processing tasks, including defocus deblurring, image denoising, super-resolution, and image enhancement. The source code and pre-trained models are available at https://github.com/swz30/MIRNetv2.

Index Terms—Multi-scale feature representation, dual-pixel defocus deblurring, image denoising, super-resolution, low-light image enhancement, and contrast enhancement

✦

1 Introduction

Owing to the physical limitations of cameras or due to complicated lighting conditions, image degradations of varying severity are often introduced as part of image acquisition. For instance, smartphone cameras come with a narrow aperture and have small sensors with limited dynamic range. Consequently, they frequently generate noisy and low-contrast images. Similarly, images captured under the unsuitable lighting are either too dark or too bright. Image restoration aims to recover the original clean image from its corrupted measurements. It is an ill-posed inverse problem, due to the existence of many possible solutions.

Recent advances in image restoration and enhancement have been led by deep learning models, as they can learn strong (generalizable) priors from large-scale datasets. Existing CNNs typically follow one of the two architecture designs: 1) an encoder-decoder, or 2) high-resolution (single-scale) feature processing. The encoder-decoder models [1], [2], [3], [4] first progressively map the input to a low-resolution representation, and then apply a gradual reverse mapping to the original resolution. Although these approaches learn a broad context by spatial-resolution reduction, on the downside, the fine spatial details are lost, making it extremely hard to recover them in the later stages. On the other hand, the high-resolution (single-scale) networks [5], [6], [7], [8] do not employ any downsampling operation, and thereby recover better spatial details. However, these networks have limited receptive field and are less effective in encoding contextual information.

Image restoration is a position-sensitive procedure, where pixel-to-pixel correspondence from the input image to the output image is needed. Therefore, it is important to remove only the undesired degraded image content, while carefully preserving the desired fine spatial details (such as true edges and texture). Such functionality for segregating the degraded content from the true signal can be better incorporated into CNNs with the help of large context, e.g., by enlarging the receptive field. Towards this goal, we

- *Syed Waqas Zamir and Aditya Arora are with the Inception Institute of Artificial Intelligence, Abu Dhabi, UAE. E-mail: waqas.zamir@inceptioniai.org, adityadvlp@gmail.com.*
- *Salman Khan and Fahad Shahbaz Khan are with the Mohammed Bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE. E-mail: salmaneme@gmail.com, fahad.khan@liu.se.*
- *Munawar Hayat is with Monash Univeristy, Melbourne, VIC 3800, Australia. E-mail: munawar.hayat@monash.edu.*
- *Ming-Hsuan Yang is with the University of California at Merced, Merced, CA 95343 USA, and also with Google, Mountain View, CA 94043 USA. E-mail: mhyang@ucmerced.edu.*
- *Ling Shao is with Terminus Group, Beijing 10000, China. E-mail: ling.shao@ieee.org.*

Manuscript received 27 Sept. 2021; revised 3 Mar. 2022; accepted 9 Apr. 2022. Date of publication 0 . 0000; date of current version 0 . 0000.
The work of Ming-Hsuan Yang was supported by NSF CAREER under Grant 1149783. The work of Ling Shao was supported by the National Natural Science Foundation of China under Grant 61929104. The work of Munawar Hayat was supported by ARC DECRA Fellowship under Grant DE200101100.
(Corresponding author: Syed Waqas Zamir.)
Recommended for acceptance by L. Zhang.
Digital Object Identifier no. 10.1109/TPAMI.2022.3167175

TABLE 1
Comparison Between MIRNet-v2 and MIRNet [9] Under the Same Experimental Settings for Image Denoising Task on the SIDD Benchmark Dataset [10]

	PSNR	Params (M)	FLOPs (B)	Convs	Activations (M)	Train Time (h)	Inference Time (ms)
MIRNet [9]	39.72	31.79	785	635	1270	139	142
MIRNet-v2 (Ours)	39.84	5.9 (81% ↓)	140 (82% ↓)	406 (36% ↓)	390 (69% ↓)	63 (55% ↓)	39 (72% ↓)

FLOPs and inference times are computed on an image of size 256×256. When compared to MIRNet [9], MIRNet-v2 is more accurate, while being significantly lighter and faster.

develop a new *multi-scale* approach that maintains the original high-resolution features along the network hierarchy, thus minimizing the loss of precise spatial details. Simultaneously, our model encodes multi-scale context by using *parallel convolution streams* that process features at lower spatial resolutions. The multi-resolution parallel branches operate in a manner that is complementary to the main high-resolution branch, thereby providing us more precise and contextually enriched feature representations.

One main distinction between our method and the existing multi-scale image processing approaches is how we aggregate contextual information. The existing methods [11], [12], [13] process each scale in isolation. In contrast, we *progressively* exchange and fuse information from coarse-to-fine resolution-levels. Furthermore, different from existing methods that employ a simple concatenation or averaging of features coming from multi-resolution branches, we introduce a new *selective kernel* fusion approach that dynamically selects the useful set of kernels from each branch representations using a self-attention mechanism. More importantly, the proposed fusion block combines features with varying receptive fields, while preserving their distinctive complementary characteristics.

The main contributions of this work include:

- A novel feature extraction model that obtains a complementary set of features across multiple spatial scales, while maintaining the original high-resolution features to preserve precise spatial details (Section 3).
- A regularly repeated mechanism for information exchange, where the features from coarse-to-fine resolution branches are progressively fused together for improved representation learning (Section 3.1).
- A new approach to fuse multi-scale features using a selective kernel network that dynamically combines variable receptive fields and faithfully preserves the original feature information at each spatial resolution (Section 3.1.1).

A preliminary version of this work has been published as a conference paper [9]. The MIRNet model [9] is expensive in terms of size and speed. In this work, we make several key modifications to MIRNet [9] that allow us to significantly reduce the computational cost while enhancing model performance (see Table 1). Specifically, in the proposed MIRNet-v2 , *(a)* We demonstrate feature fusion only in the direction from low- to high-resolution streams performs best, and the information flow from high- to low-resolution branches can be removed to improve efficiency. *(b)* We replace the dual attention unit with a new residual contextual block (RCB). Furthermore, we introduce group convolutions in RCB that are capable of learning unique representations in each filter group, while being more resource efficient than standard convolutions. *(c)* We employ progressive learning to improve training speed: the network is trained on small image patches in the early epochs and on gradually large patches in the later training epochs. *(d)* We show the effectiveness of the proposed design on a new task of dual-pixel defocus deblurring [14] alongside the other image processing tasks of image denoising, super-resolution and image enhancement. Our MIRNet-v2 achieves state-of-the-results on *all* six datasets. Furthermore, we extensively evaluate our approach on practical challenges, such as generalization ability across datasets (Section 4)

In Table 1, we compare MIRNet-v2 with MIRNet [9] under the same training and inference settings. The results show that MIRNet-v2 is more accurate (improving PSNR from 39.72 dB to 39.84 dB), while reducing the number of parameters and FLOPs by $\sim 81\%$, convolutions by 36%, and activations by 69%. Furthermore, the training and inference speed is increased by $2.2\times$ and $3.6\ \times$, respectively.

2 Related Work

Rapidly growing image content necessitates the need to develop effective image restoration and enhancement algorithms. In this paper, we propose a new method capable of performing dual-pixel defocus deblurring, image denoising, super-resolution, and image enhancement. Unlike existing works for these problems, our approach processes features at the original resolution in order to preserve spatial details, while effectively fuses contextual information from multiple parallel branches. Next, we briefly describe the representative methods for each of the studied problems.

2.1 Dual-Pixel Defocus Deblurring

Images captured with wide camera aperture have shallow depth of field (DoF), where the scene regions that lie outside the DoF are out-of-focus. Given an image with defocus blur, the goal of defocus deblurring is to generate an all-in-focus image. Existing defocus deblurring approaches either directly deblur images [14], [15], [16], or first estimate the defocus dispartiy map and then use it to guide the deblurring procedure [17], [18], [19]. Modern cameras are equipped with dual-pixel sensor that has two photodiodes at each pixel location, thereby generating two sub-aperture views. The phase difference between these views is useful in measuring the amount of defocus blur at each scene point. Recently, Abuolaim *et al.* [14] presented a dual-pixel deblurring dataset (DPDD) and a new method based on encoder-decoder design. In this paper our focus is also on

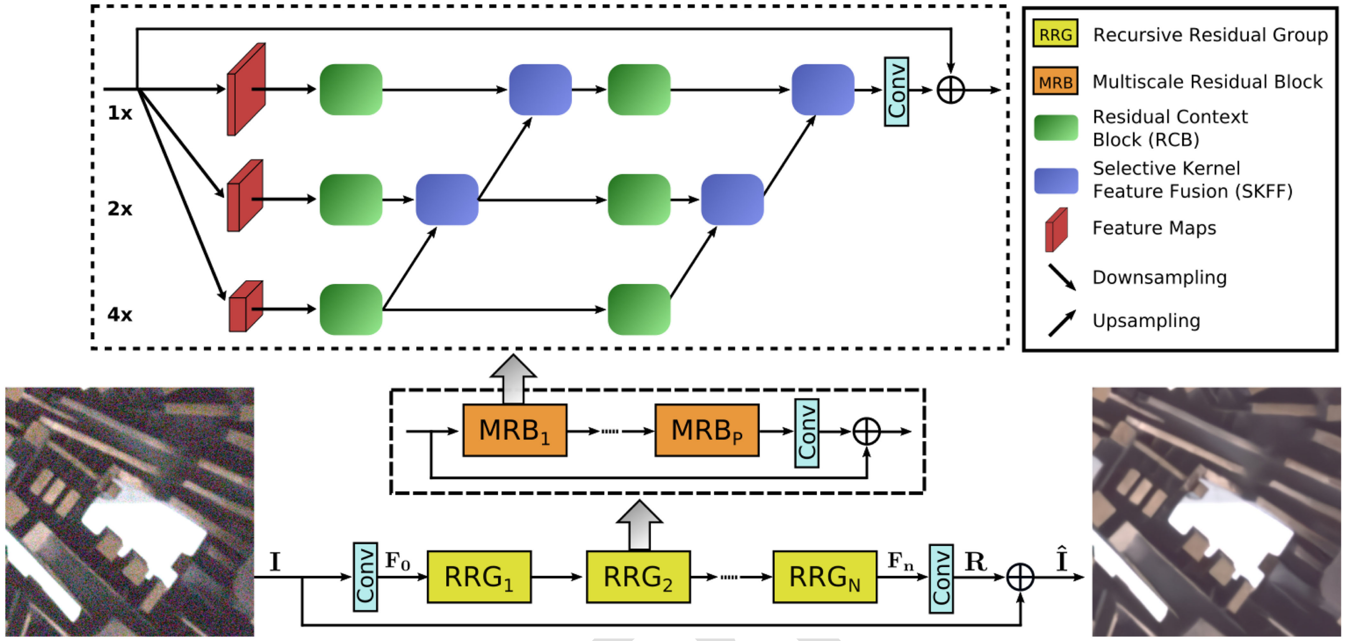

1x
2x
4x
Conv
RRG Recursive Residual Group
MRB Multiscale Residual Block
Residual Context Block (RCB)
Selective Kernel Feature Fusion (SKFF)
Feature Maps
Downsampling
Upsampling
MRB$_1$
MRB$_P$
Conv
I
Conv
F$_0$
RRG$_1$
RRG$_2$
RRG$_N$
F$_n$
Conv
R
Î

Fig. 1. Framework of the proposed MIRNet-v2 that learns enriched feature representations for image restoration and enhancement. MIRNet-v2 is based on a recursive residual design. In the core of MIRNet-v2 is the multi-scale residual block (MRB) whose main branch is dedicated to maintaining spatially-precise high-resolution representations through the entire network and the complimentary set of parallel branches provide better contextualized features.

deblurring images directly using the dual-pixel data as in [14], [16]. Previous defocus deblurring works [14], [16] employ the encoder-decoder that repeatedly uses the downsampling operation, thus causing significant fine detail loss. Whereas the architectural design of our approach enables preservation of desired textural details in the restored image.

2.2 Image Denoising

Classic denoising methods are mainly based on modifying transform coefficients [20], [21] or averaging neighborhood pixels [22], [23], [24]. Although the classical approaches perform well, the self-similarity [25] based algorithms, e.g., NLM [26] and BM3D [27], demonstrate promising denoising performance. Numerous patch-based schemes that exploit redundancy (self-similarity) in images are later developed [28], [29], [30], [31]. Recently, deep learning models [6], [9], [32], [33], [34], [35], [36], [37], [38], [39], [40], [41], [42] make significant advances in image denoising, yielding favorable results than those of the hand-crafted methods.

2.3 Image Super-Resolution

Prior to the deep-learning era, numerous super-resolution (SR) algorithms have been proposed based on the sampling theory [43], [44], edge-guided interpolation [45], [46], natural image priors [47], [48], patch-exemplars [49], [50] and sparse representations [51], [52]. Currently, deep-learning techniques are being actively explored as they provide dramatically improved results over conventional algorithms. The data-driven SR approaches differ according to their architecture designs [53], [54], [55]. Early methods [5], [56] take a low-resolution (LR) image as input and learn to directly generate its high-resolution (HR) version. In contrast to directly producing a latent HR image, recent SR networks [57], [58], [59], [60] employ the residual learning framework [61] to learn the high-frequency image detail, which is later added to the input LR image to produce the final result. Other networks designed to perform SR include recursive learning [62], [63], [64], progressive reconstruction [65], [66], dense connections [7], [67], [68], attention mechanisms [69], [70], [71], multi-branch learning [66], [72], [73], [74], and generative adversarial networks (GANs) [68], [75], [76], [77].

2.4 Image Enhancement

Oftentimes, cameras generate images that lack vivid details or contrast. A number of factors contribute to the low quality of images, including unsuitable lighting conditions and physical limitations of camera devices. For image enhancement, histogram equalization is the most commonly used approach. However, it frequently produces under- or over-enhanced images. Motivated by the Retinex theory [78], several enhancement algorithms mimicking human vision have been proposed in the literature [79], [80], [81], [82]. Recently, CNNs have been successfully applied to general, as well as low-light, image enhancement problems [83]. Notable works employ Retinex-inspired networks [4], [84], [85], [86], encoder-decoder networks [87], [88], [89], [90], [91], and GANs [92], [93], [94].

3 Proposed Method

A schematic of the proposed MIRNet-v2 is shown in Fig. 1. We first present an overview of the proposed MIRNet-v2 for image restoration and enhancement. We then provide details of the *multi-scale residual block*, which is the fundamental building block of our method, containing several key elements: *(a)* parallel multi-resolution convolution streams for extracting (fine-to-coarse) semantically-richer and (coarse-to-

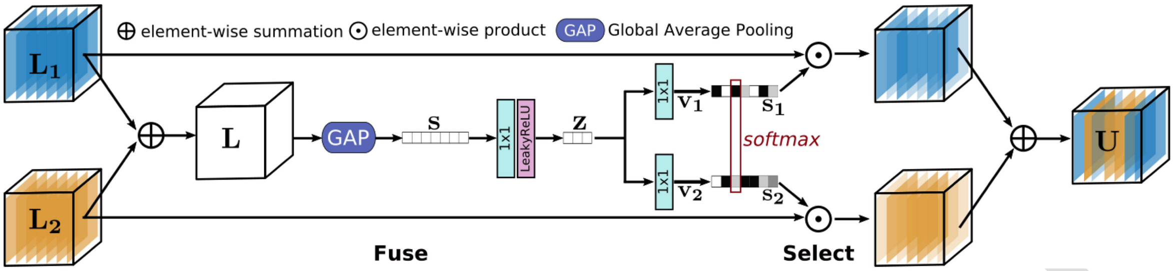


Fig. 2. Schematic for selective kernel feature fusion (SKFF). It operates on features from different resolution streams, and performs aggregation based on self-attention.

fine) spatially-precise feature representations, *(b)* information exchange across multi-resolution streams, *(c)* attention-based aggregation of features arriving from different streams, and *(d)* residual contextual blocks to extract attention-based features.

Overall Pipeline. Given an image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$, the proposed model first applies a convolutional layer to extract low-level features $\mathbf{F_0} \in \mathbb{R}^{H\times W\times C}$. Next, the feature maps $\mathbf{F_0}$ pass through N number of recursive residual groups (RRGs), yielding deep features $\mathbf{F_n} \in \mathbb{R}^{H\times W\times C}$. We note that each RRG contains several multi-scale residual blocks, which is described in Section 3.1. Next, we apply a convolution layer to deep features $\mathbf{F_n}$ and obtain a residual image $\mathbf{R} \in \mathbb{R}^{H\times W\times 3}$. Finally, the restored image is obtained as $\hat{\mathbf{I}} = \mathbf{I} + \mathbf{R}$. We optimize the proposed network using the Charbonnier loss [95]

$$\mathcal{L}(\hat{\mathbf{I}}, \mathbf{I}^*) = \sqrt{\|\hat{\mathbf{I}} - \mathbf{I}^*\|^2 + \varepsilon^2}, \qquad (1)$$

where $\mathbf{I}^*$ denotes the ground-truth image, and ε is a constant which we empirically set to 10^{-3} for all the experiments.

3.1 Multi-Scale Residual Block

To encode context, existing CNNs [1], [96], [97], [98], [99], [100] typically employ the following architecture design: *(a)* the receptive field of neurons is fixed in *each* layer/stage, *(b)* the spatial size of feature maps is *gradually* reduced to generate a semantically strong low-resolution representation, and *(c)* a high-resolution representation is *gradually* recovered from the low-resolution representation. However, it is well-understood in vision science that in the primate visual cortex, the sizes of the local receptive fields of neurons in the same region are different [101], [102], [103], [104]. Therefore, a similar mechanism of collecting multi-scale spatial information in the same layer is more effective when incorporated with in CNNs [105], [106], [107], [108]. Motivated by this, we propose the multi-scale residual block (MRB), as shown in Fig. 1. It is capable of generating a spatially-precise output by maintaining high-resolution representations, while receiving rich contextual information from low-resolutions. The MRB consists of multiple (three in this paper) fully-convolutional streams connected in parallel that operate on varying resolution feature maps (ranging from low to high). It allows contextualized-information transfer from the low-resolution streams to consolidate the high-resolution features. Next, we describe the individual components of MRB.

3.1.1 Selective Kernel Feature Fusion

One fundamental property of neurons present in the visual cortex is their ability to change receptive fields according to the stimulus [109]. This mechanism of adaptively adjusting receptive fields can be incorporated in CNNs by using multi-scale feature generation (in the same layer) followed by feature aggregation and selection. The most commonly used approaches for feature aggregation include simple concatenation or summation. However, these choices provide limited expressive power to the network, as reported in [109]. In MRB, we introduce a nonlinear procedure for fusing features coming from different resolution streams using a self-attention mechanism. Motivated by [109], we call it selective kernel feature fusion (SKFF).

The SKFF module performs dynamic adjustment of receptive fields via two operations – *Fuse* and *Select*, as illustrated in Fig. 2. The *fuse* operator generates global feature descriptors by combining the information from multi-resolution streams. The *select* operator uses these descriptors to recalibrate the feature maps (of different streams) followed by their aggregation. Next, we provide details of both operators. *(1) Fuse:* SKFF receives inputs from two parallel convolution streams carrying different scales of information. We first combine these multi-scale features using an element-wise sum as: $\mathbf{L} = \mathbf{L_1} + \mathbf{L_2}$. We then apply global average pooling (GAP) across the spatial dimension of $\mathbf{L} \in \mathbb{R}^{H\times W\times C}$ to compute channel-wise statistics $\mathbf{s} \in \mathbb{R}^{1\times 1\times C}$. Next, we apply a channel-downscaling convolution layer to generate a compact feature representation $\mathbf{z} \in \mathbb{R}^{1\times 1\times r}$, where $r = \frac{C}{8}$ for all our experiments. Finally, the feature vector $\mathbf{z}$ passes through two parallel channel-upscaling convolution layers (one for each resolution stream) and provides us with two feature descriptors $\mathbf{v_1}$ and $\mathbf{v_2}$, each with dimensions $1 \times 1 \times C$. *(2) Select:* This operator applies the softmax function to $\mathbf{v_1}$ and $\mathbf{v_2}$, yielding attention activations $\mathbf{s_1}$ and $\mathbf{s_2}$ that we use to adaptively recalibrate multi-scale feature maps $\mathbf{L_1}$ and $\mathbf{L_2}$, respectively. The overall process of feature recalibration and aggregation is defined as: $\mathbf{U} = \mathbf{s_1} \cdot L_1 + s_2 \cdot L_2$. Note that the SKFF uses ~ 5x fewer parameters than aggregation with concatenation but generates more favorable results (an ablation study is provided in the experiments section).

3.1.2 Residual Contextual Block

While the SKFF block fuses information across multi-resolution branches, we also need a distillation mechanism to extract useful information from within a feature tensor. Motivated by the advances of recent low-level vision

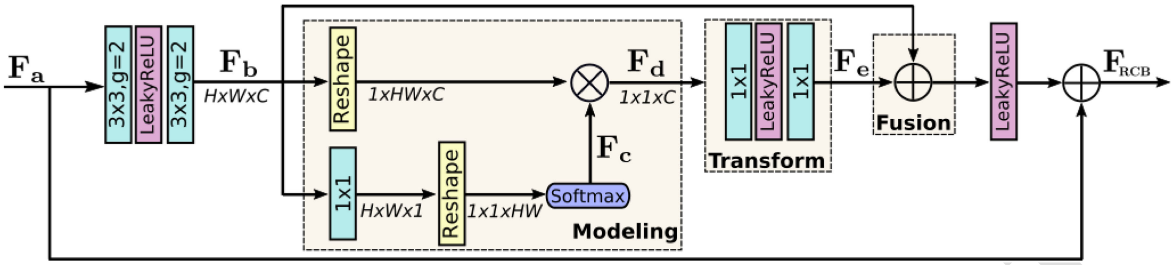


Fig. 3. Architecture of residual contextual block (RCB). In the first two group convolution layers, g represents the number of groups. $\otimes$ denotes matrix multiplication.

methods [32], [69], [70], [71] which incorporate attention mechanisms [110], [111], [112], we propose the residual contextual block (RCB) to extract features in the convolutional streams. The schematic of RCB is shown in Fig. 3. The RCB suppresses less useful features and only allows more informative ones to pass further. The overall process of RCB is summarized as

$$\mathbf{F_{RCB}} = \mathbf{F_a} + W(\mathrm{CM}(\mathbf{F_b})), \tag{2}$$

where $\mathbf{F_b} \in \mathbb{R}^{H\times W\times C}$ represents feature maps that are obtained by applying two 3x3 *group* convolution layers to the input features $\mathbf{F_b} \in \mathbb{R}^{H\times W\times C}$ at the beginning of the RCB. These group convolutions are more resource efficient than standard convolutions and capable of learning unique representations in each filter group. W denotes the last convolutional layer with filter size 1x1. CM stands for contextual module that is realized in three parts. *(1) Context modeling:* From the original feature maps $\mathbf{F_b}$, we first generate new features $\mathbf{F_c} \in \mathbb{R}^{1\times 1\times HW}$ by applying 1x1 convolution followed by the reshaping and softmax operations. Next we reshape $\mathbf{F_b}$ to $\mathbb{R}^{1\times HW\times C}$ and perform matrix multiplication with $\mathbf{F_c}$ to obtain the global feature descriptor $\mathbf{F_d} \in \mathbb{R}^{1\times 1\times C}$. *(2) Feature transform:* To capture the inter-channel dependencies we pass the descriptor $\mathbf{F_d}$ through two 1x1 convolutions, resulting in new attention features $\mathbf{F_e} \in \mathbb{R}^{1\times 1\times C}$. *(3) Feature fusion:* We employ element-wise addition operation to aggregate contextual features $\mathbf{F_e}$ to each position of the original features $\mathbf{F_b}$.

3.2 Progressive Training Regime

When considering the image patch size for network training, there is a trade-off between the training speed and test-time accuracy [113], [114]. On large patches, CNNs capture fine image details to provide improved results, but they are slower to train. Whereas, training on small image patches is faster, but comes at the cost of accuracy drop. To strike the right balance between the training speed and accuracy, we propose a progressive learning method where the network is trained on smaller image patches in the early epochs and on gradually larger patches in the later training epochs. This approach can also be understood as a curriculum learning process where the network sequentially moves from learning a simpler task to a more complex one (where modeling of fine details is required). The progressive learning strategy on mixed-size image patches not only improves the training speed but also enhances the model performance at test time where the input images can be of different sizes (which is common in image restoration problems).

4 Experiments

In this section, we perform qualitative and quantitative assessments of the results produced by our MIRNet-v2 and compare it with the state-of-the-art methods. Next, we describe the datasets, and then provide the implementation details. Finally, we report results for *(a)* dual-pixel defocus deblurring, *(b)* image denoising, *(c)* image super-resolution and *(d)* image enhancement, on six real image datasets.

4.1 Real Image Datasets

Dual-Pixel Defocus Deblurring. DPDD [14] dataset contains 500 indoor/outdoor scenes captured with a DSLR camera. Each scene consists of two defocus blurred sub-aperture views captured with a wide camera aperture, and the corresponding all-in-focus ground truth image captured with a narrow aperture. The DDPD dataset is divided into 350 images for training, 74 images for validation and 76 images for testing.

Image Denoising. (1) DND [115] consists of 50 images captured with four consumer cameras. Since the images are of very high-resolution, the dataset providers extract 20 crops of size 512×512 from each image, yielding 1000 patches in total. All these patches are used for testing (as DND does not contain training or validation sets). The ground-truth noise-free images are not released publicly, therefore the image quality scores in terms of PSNR and SSIM can only be obtained through an online server [116].

(2) SIDD [10] is collected with smartphone cameras. Due to the small sensor and high-resolution, the noise levels in smartphone images are much higher than those of DSLRs. SIDD contains 320 image pairs for training and 1280 for validation.

Super-Resolution. RealSR [117] contains real-world LR-HR image pairs of the same scene captured by adjusting the focal-length of the cameras. RealSR has both indoor and outdoor images taken with two cameras. The number of training image pairs for scale factors ×2, ×3 and ×4 are 183, 234 and 178, respectively. For each scale factor, 30 test images are also provided in RealSR.

Image Enhancement. (1) LoL [85] is created for low-light image enhancement problem. It provides 485 images for training and 15 for testing. Each image pair in LoL consists of a low-light input image and its corresponding well-exposed reference image.

(2) MIT-Adobe FiveK [118] contains 5000 images of various indoor and outdoor scenes captured with DSLR cameras in different lighting conditions. The tonal attributes

TABLE 2
Dual-Pixel Defocus Deblurring Comparisons on the DPDD Dataset [14]

	Indoor Scenes				Outdoor Scenes				Combined			
Method	PSNR ↑	SSIM ↑	MAE ↓	LPIPS ↓	PSNR ↑	SSIM ↑	MAE ↓	LPIPS ↓	PSNR ↑	SSIM ↑	MAE ↓	LPIPS ↓
EBDB [17]	25.77	0.772	0.040	0.297	21.25	0.599	0.058	0.373	23.45	0.683	0.049	0.336
DMENet [19]	25.50	0.788	0.038	0.298	21.43	0.644	0.063	0.397	23.41	0.714	0.051	0.349
JNB [18]	26.73	0.828	0.031	0.273	21.10	0.608	0.064	0.355	23.84	0.715	0.048	0.315
DPDNet [14]	27.48	0.849	0.029	0.189	22.90	0.726	0.052	0.255	25.13	0.786	0.041	0.223
RDPD [16]	28.10	0.843	0.027	0.210	22.82	0.704	0.053	0.298	25.39	0.772	0.040	0.255
MIRNet-v2 (Ours)	**28.96**	**0.881**	**0.024**	**0.154**	**23.59**	**0.753**	**0.049**	**0.205**	**26.20**	**0.816**	**0.037**	**0.180**

The test set of DPDD contains 37 indoor scenes and 39 outdoor scenes. Best and second best scores are highlighted *and* underlined, *respectively.*

of all images are manually adjusted by five different trained photographers (labelled as experts A to E). Similar to [119], [120], [121], we also consider the enhanced images of expert C as the ground-truth. Moreover, the first 4500 images are used for training and the last 500 for testing.

4.2 Implementation Details

The proposed architecture is end-to-end trainable and requires no pre-training of sub-modules. We train four different networks for four different restoration tasks. For the dual-pixel defocus deblurring, we concatenate the left and right sub-aperture images and feed them as input to the

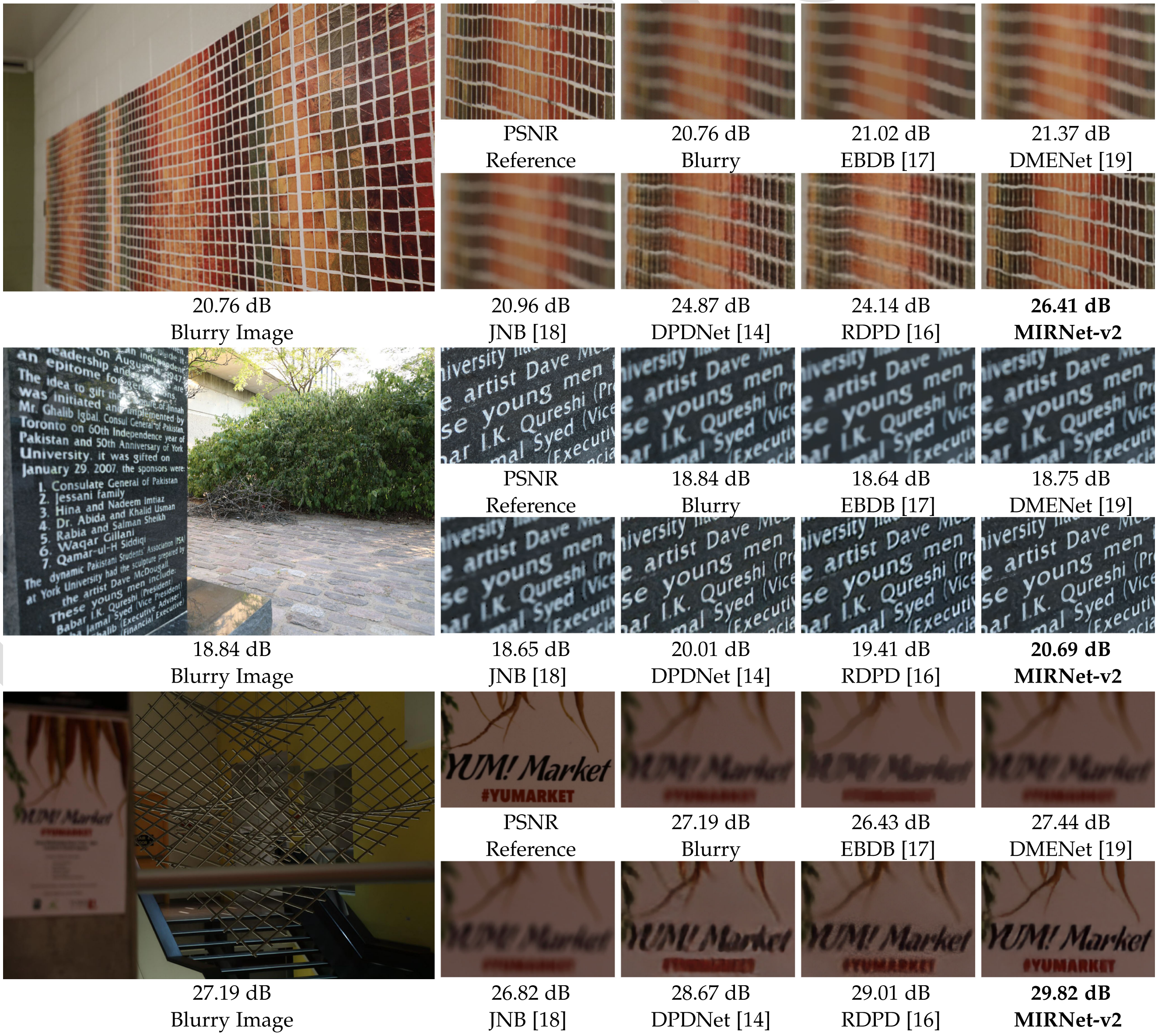


Fig. 4. Visual comparisons for dual-pixel defocus deblurring on the DPDD dataset [14]. Compared to the other approaches, our MIRNet-v2 more effectively removes blur while preserving the fine image details.

TABLE 3
Denoising Comparisons on SIDD [10] and DND [115] Datasets

	SIDD [10]		DND [115]	
Method	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
DnCNN [6]	23.66	0.583	32.43	0.790
MLP [123]	24.71	0.641	34.23	0.833
BM3D [27]	25.65	0.685	34.51	0.851
CBDNet* [34]	30.78	0.801	38.06	0.942
DAGL [124]	38.94	0.953	39.77	0.956
RIDNet* [32]	38.71	0.951	39.26	0.953
AINDNet* [41]	38.95	0.952	39.37	0.951
VDN [40]	39.28	0.956	39.38	0.952
DeamNet* [125]	39.47	0.957	39.63	0.953
SADNet* [38]	39.46	0.957	39.59	0.952
DANet+* [39]	39.47	0.957	39.58	0.955
CycleISP* [37]	39.52	0.957	39.56	**0.956**
MIRNet-v2 (Ours)	**39.84**	**0.959**	**39.86**	0.955

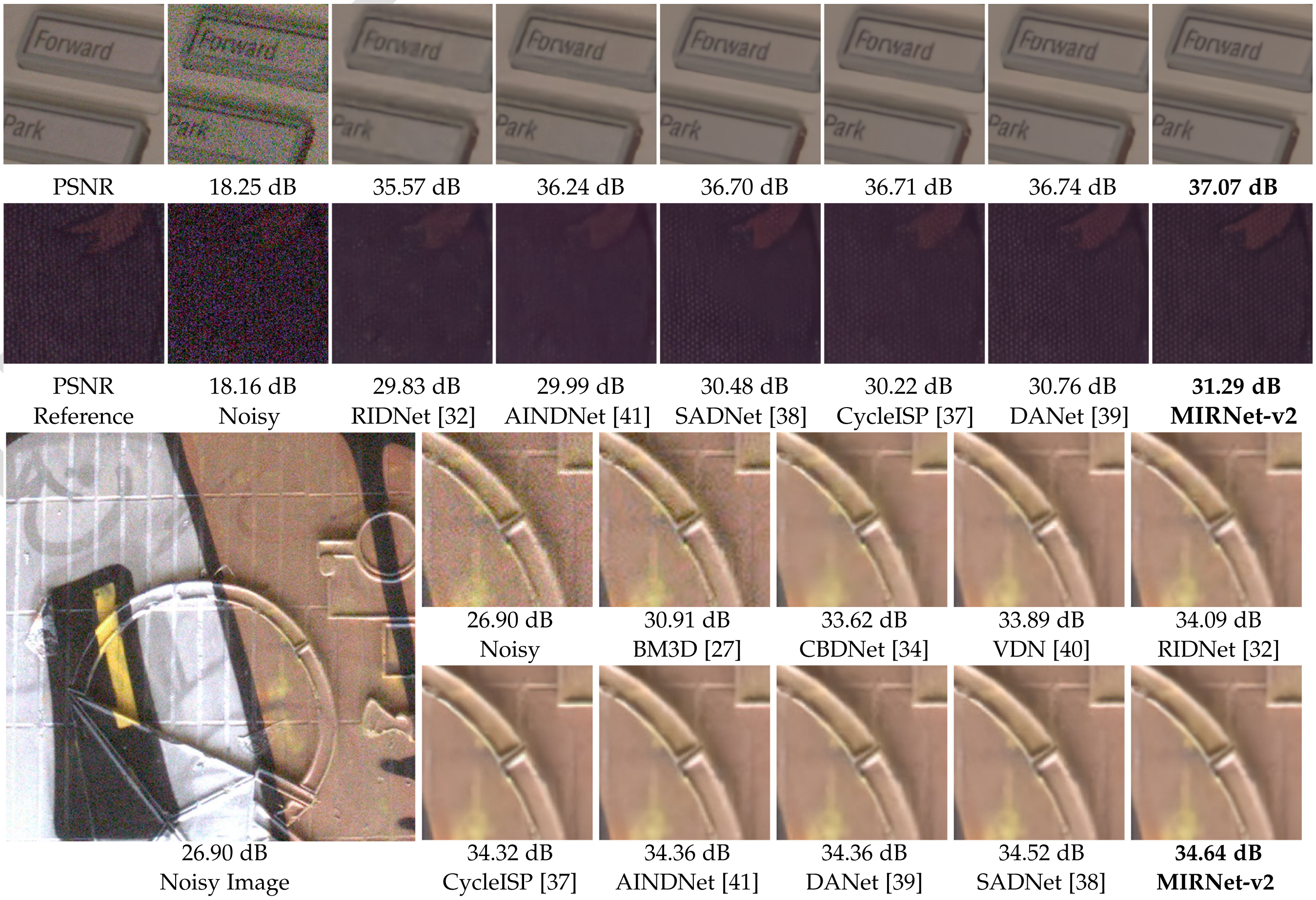
$*$ indicates the methods that use additional training data. Whereas our MIRNet-v2 is only trained on the SIDD iand directly tested on DND.

network. The training parameters, common to all experiments, are the following. We use 4 RRGs, each of which further contains 2 MRBs. The MRB has 3 parallel streams with channel dimensions of $80, 120, 180$ at resolutions $1, \frac{1}{2}, \frac{1}{4}$, respectively. Each stream in MRB has 2 RCBs with shared parameters. The models are trained with the Adam optimizer ($\beta_1 = 0.9$, and $\beta_2 = 0.999$) for 3×10^5 iterations. The initial learning rate is set to 2×10^{-4}. We employ the cosine annealing strategy [122] to steadily decrease the learning rate from initial value to 10^{-6} during training. For progressive training, we use the image patch sizes of 128, 144, 192, and 224. The batch size is set to 64 and, for data augmentation, we perform horizontal and vertical flips.

4.3 Dual-Pixel Defocus Deblurring

We compare the performance of the proposed MIRNet-v2 with the conventional defocus deblurring methods (EBDB [17] and JNB [18]) as well as the learning-based approaches (DMENet [19], DPDNet [14], and RDPD [16]). Table 2 shows that our method achieves state-of-the-art results for both the indoor and outdoor scene categories. In particular, our MIRNet-v2 achieves 0.86 dB PSNR improvement over the previous best method RDPD [16] on indoor images and 0.77 dB on outdoor images. When both scene categories are combined, our method shows performance gains of 0.81 dB over RDPD [14] and 1.07 dB over the second best method DPDNet [14].

In Fig. 4, we provide defocus-deblurred results produced by different methods for both indoor and outdoor scenes. It is noticeable that our method effectively removes the spatially varying defocus blur and produces images that are more sharper and visually faithful to the ground-truth than those of the compared approaches.



PSNR 18.25 dB 35.57 dB 36.24 dB 36.70 dB 36.71 dB 36.74 dB **37.07 dB**

PSNR 18.16 dB 29.83 dB 29.99 dB 30.48 dB 30.22 dB 30.76 dB **31.29 dB**

Reference Noisy RIDNet [32] AINDNet [41] SADNet [38] CycleISP [37] DANet [39] **MIRNet-v2**

26.90 dB 30.91 dB 33.62 dB 33.89 dB 34.09 dB

Noisy BM3D [27] CBDNet [34] VDN [40] RIDNet [32]

26.90 dB 34.32 dB 34.36 dB 34.36 dB 34.52 dB **34.64 dB**

Noisy Image CycleISP [37] AINDNet [41] DANet [39] SADNet [38] **MIRNet-v2**

Fig. 5. Image denoising comparisons. First two examples are from SIDD [10] and the last is from DND [115]. The proposed MIRNet-v2 better preserves fine texture and structural patterns in the denoised images.

TABLE 4
Super-Resolution Evaluation on the RealSR Dataset [117]

Scale	x2		x3		x4	
Method	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Bicubic	32.61	0.907	29.34	0.841	27.99	0.806
VDSR [57]	33.64	0.917	30.14	0.856	28.63	0.821
SRResNet [77]	33.69	0.919	30.18	0.859	28.67	0.824
RCAN [69]	33.87	0.922	30.40	0.862	28.88	0.826
LP-KPN [117]	33.90	0.927	30.42	0.868	28.92	0.834
MIRNet-v2 (Ours)	**34.38**	**0.934**	**31.15**	**0.883**	**29.16**	**0.845**

Compared to the state-of-the-art, our method consistently yields significantly better image quality scores for all three scaling factors.

4.4 Image Denoising

In this section, we demonstrate the effectiveness of the proposed MIRNet-v2 for image denoising. We train our network only on the training set of the SIDD [10] and directly evaluate it on the test images of both SIDD and DND [115] datasets. Quantitative comparisons in terms of PSNR and SSIM metrics are summarized in Table 3. Our MIRNet-v2 performs favourably against the data-driven, as well as conventional, denoising algorithms. Specifically, when compared to the recent best methods, our algorithm demonstrates a performance gain of 0.32 dB over CycleISP [37] on SIDD and 0.11 dB over DAGL [124] on DND. Furthermore, it is worth noting that CycleISP [37] uses additional training data, yet our method yields considerably better results.

Fig. 5 shows a visual comparisons of our results with those of other competing algorithms. The MIRNet-v2 is effective in removing real noise and produces perceptually-pleasing and sharp images. Moreover, it is can maintain the spatial smoothness of the homogeneous regions without introducing artifacts. In contrast, most of the other methods either yield over-smooth images and thus sacrifice structural content and fine textural details, or produce images with chroma artifacts and blotchy texture.

Generalization Capability. The DND and SIDD datasets are acquired with different sets of cameras having different noise characteristics. Since the DND benchmark does not provide training data, setting a new state-of-the-art on DND

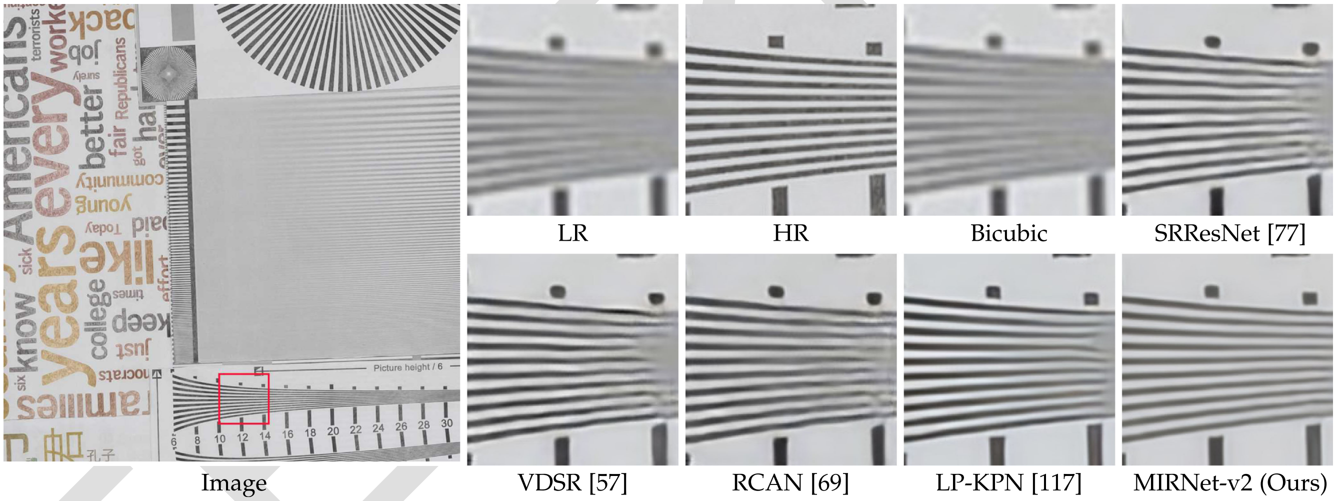


Fig. 6. Comparisons for ×4 super-resolution on the RealSR [117] dataset. The image produced by our MIRNet-v2 is more faithful to the ground-truth than other competing methods (see lines near the right edge of the crops).

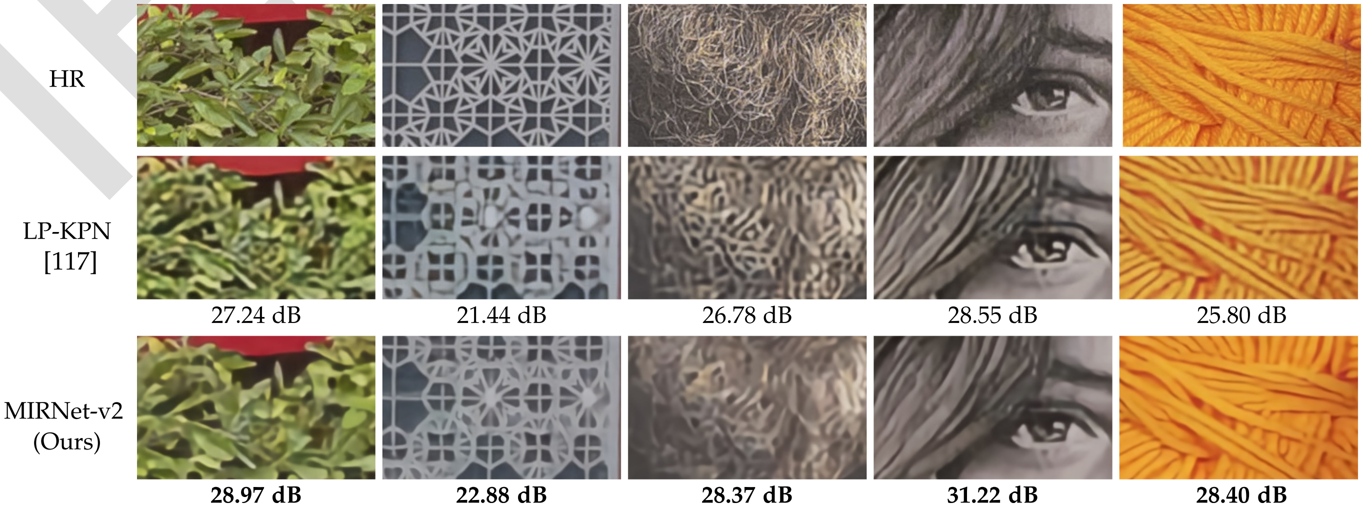


Fig. 7. Additional visual examples for ×4 super-resolution, comparing our MIRNet-v2 against the state-of-the-art approach [117]. Note that all example crops are taken from different images.

TABLE 5
Low-Light Image Enhancement Evaluation on the LoL Dataset [85]

Method	BIMEF	CRM	Dong	LIME	MF	RRM	SRIE	Retinex-Net	MSR	NPE	GLAD	KinD	KinD++	MIRNet-v2
	[126]	[127]	[128]	[129]	[130]	[131]	[130]	[85]	[81]	[132]	[133]	[4]	[134]	(Ours)
PSNR	13.86	17.20	16.72	16.76	18.79	13.88	11.86	16.77	13.17	16.97	19.72	20.87	<u>21.30</u>	**24.74**
SSIM	0.577	0.644	0.582	0.564	0.642	0.658	0.498	0.559	0.479	0.589	0.703	0.810	<u>0.822</u>	**0.851**

The proposed method significantly advances the state-of-the-art.

TABLE 6
Image Enhancement Comparisons on the MIT-Adobe FiveK Dataset [118]

Method	HDRNet [135]	W-Box [119]	DR [120]	DPE [92]	DeepUPE [121]	MIRNet-v2 (Ours)
PSNR	21.96	18.57	20.97	22.15	<u>23.04</u>	**23.97**
SSIM	0.866	0.701	0.841	0.850	<u>0.893</u>	**0.931**

with our SIDD trained network indicates the good generalization capability of our approach.

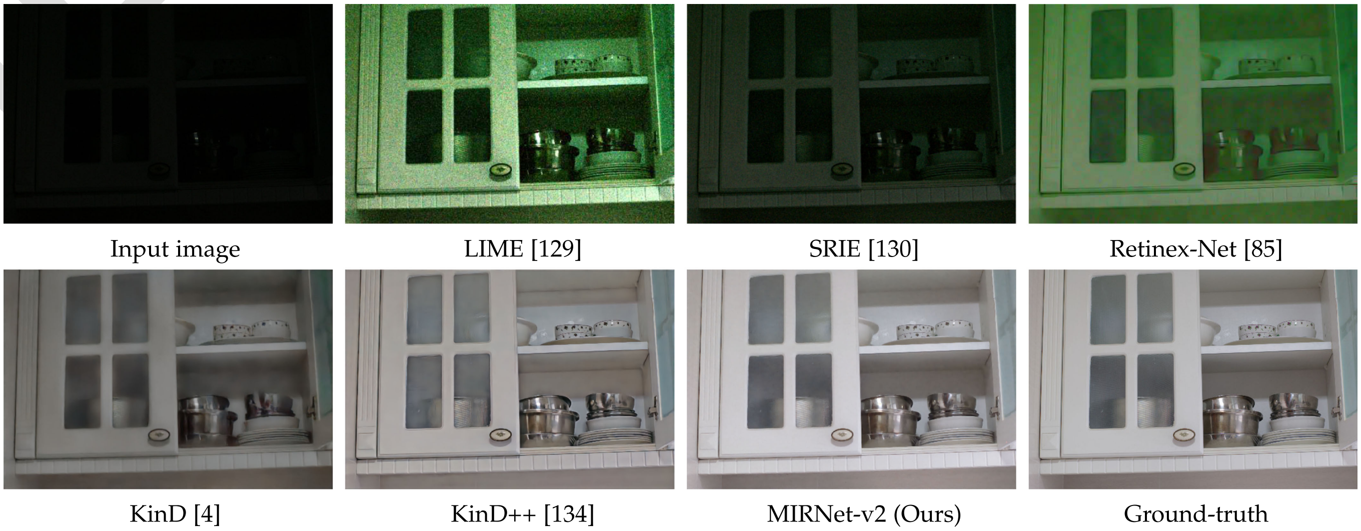
4.5 Super-Resolution

We compare our MIRNet-v2 against the state-of-the-art SR algorithms (VDSR [57], SRResNet [77], RCAN [69], LP-KPN [117]) on the testing images of the RealSR [117] for upscaling factors of $\times 2$, $\times 3$ and $\times 4$. Note that all the benchmarked algorithms are trained on the RealSR [117] dataset for a fair comparison. In the experiments, we also include bicubic interpolation [43], which is the most commonly used method for generating super-resolved images. Here, we compute the PSNR and SSIM scores using the Y channel (in YCbCr color space), as it is a common practice in the SR literature [53], [54], [69], [117]. The results in Table 4 show that the bicubic interpolation provides the least accurate results, thereby indicating its low suitability for dealing with real images. Moreover, the same table shows that the recent method LP-KPN [117] achieves marginal improvement of only ~ 0.04 dB over the previous best method RCAN [69]. In contrast, our method significantly advances state-of-the-art and consistently achieves better image quality scores than other approaches for all three scaling factors. Particularly, compared to LP-KPN [117], our method leads to performance gains of 0.48 dB, 0.73 dB, and 0.24 dB for scaling factors $\times 2$, $\times 3$ and $\times 4$, respectively. The trend is similar for the SSIM metric as well.

Visual comparisons in Fig. 6 show that our MIRNet-v2 can effectively recover content structures. In contrast, VDSR [57], SRResNet [77] and RCAN [69] reproduce results with noticeable artifacts. Furthermore, LP-KPN [117] is not able to preserve structures (see near the right edge of the crop). Several more examples are provided in Fig. 7 to further compare the image reproduction quality of our method against the previous best method [117]. It can be seen that LP-KPN [117] has a tendency to over-enhance the contrast (cols. 1, 3, 4) and in turn causes loss of details near dark and high-light areas. In contrast, the proposed MIRNet-v2 successfully reconstructs structural patterns and edges (col. 2) and produces images that are natural (cols. 1, 4) and have better color reproduction (col. 5).

4.6 Image Enhancement

In this section, we demonstrate the effectiveness of our algorithm by evaluating it for the image enhancement task. We report PSNR/SSIM values of our method and several other



Input image

LIME [129]

SRIE [130]

Retinex-Net [85]

KinD [4]

KinD++ [134]

MIRNet-v2 (Ours)

Ground-truth

Fig. 8. Visual comparison of low-light enhancement approaches on the LoL dataset [85]. The image produced by our method is visually closer to the ground-truth in terms of brightness and global contrast.

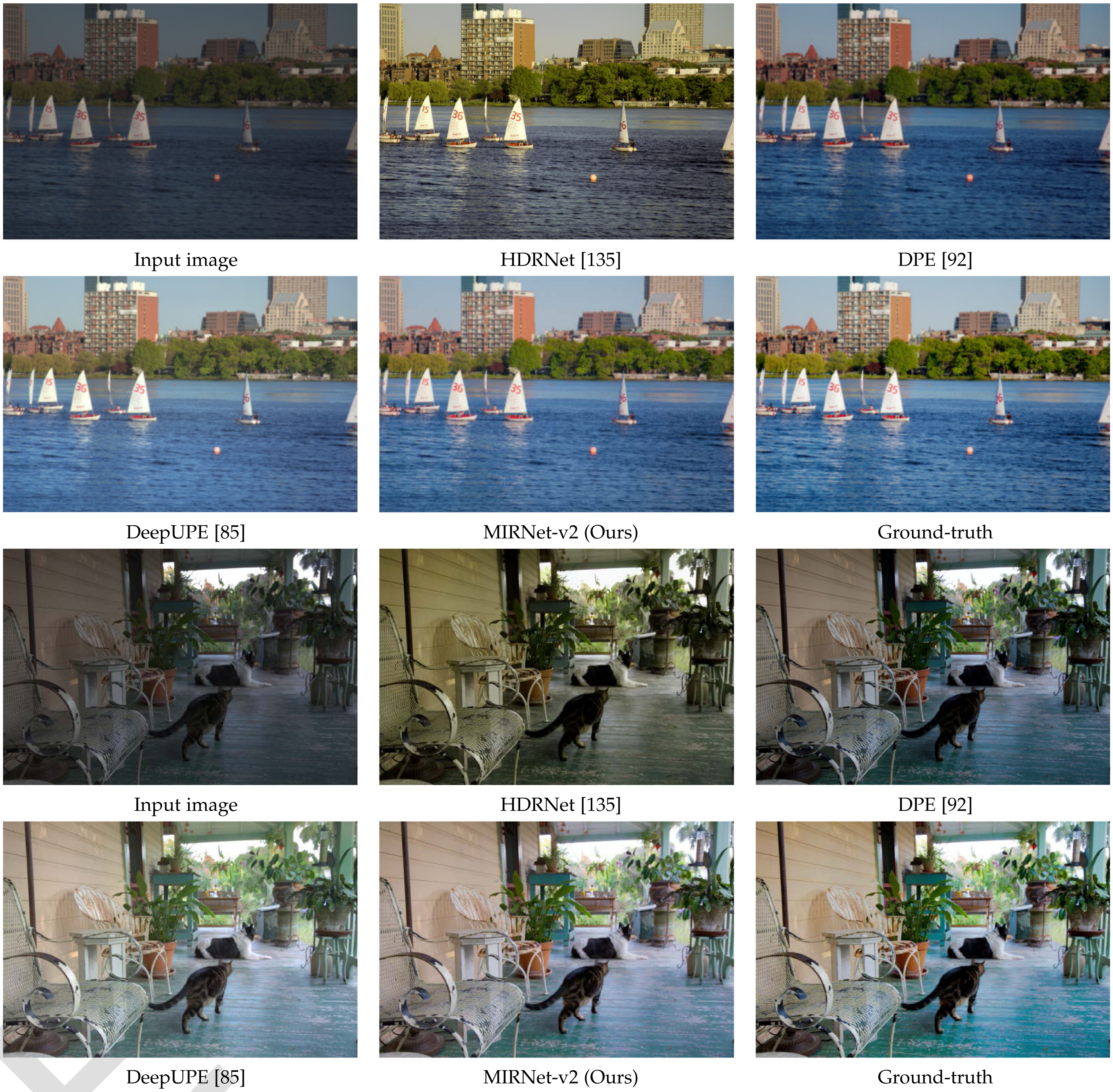


Fig. 9. Visual results of image enhancement on the MIT-Adobe FiveK [118] dataset. Compared to the state-of-the-art, our MIRNet-v2 makes better color and contrast adjustments and produces images that appear vivid, natural and pleasant.

techniques in Tables 5 and 6 for the LoL [85] and MIT-Adobe FiveK [118] datasets, respectively. It can be seen that our MIRNet-v2 achieves significant improvements over previous approaches. Notably, when compared to the recent best methods, MIRNet-v2 obtains 3.44 dB performance gain over KinD++ [134] on the LoL dataset and 0.93 dB improvement over DeepUPE[1] [121] on the Adobe-Fivek dataset.

We show visual results in Figs. 8 and 9. Compared to other techniques, our method generates enhanced images that are natural and vivid in appearance and have better global and local contrast.

4.7 Ablation Studies

We study the impact of each of our architectural components and design choices on the final performance. All the ablation experiments are performed for the super-resolution task with ×3 scale factor. The ablation models are trained on image patches of size 128×128 for 10^5 iterations. Table 7 shows that removing skip connections causes the largest performance drop. Without skip connections, the network finds it difficult to converge and yields high training errors, and consequently low PSNR. Furthermore, the information exchange among parallel convolution streams via SKFF is helpful and leads to improved performance. Similarly, RCB contributes positively towards the final image quality.

Table 8 shows that the proposed RCB provides favorable performance gain over the baseline Resblock from

[1] Note that the quantitative results reported in [121] are incorrect. The correct scores are later released by the original authors [link].

TABLE 7
Impact of Individual Components of MRB

Skip connections		✓	✓	✓	✓
RCB	✓		✓		✓
SKFF intermediate	✓	✓			✓
SKFF final	✓	✓	✓	✓	✓
PSNR (in dB)	28.21	30.79	30.85	30.68	**30.97**

TABLE 8
Effect of Individual Components of RCB

	PSNR	Params (M)	FLOPs (B)
Baseline [72], g=2	30.84	5.0	139.5
+ RCB, g=2	30.97	5.9	139.8
RCB w/o transform, g=2	30.92	5.0	139.7
RCB, g=1	31.05	9.7	253.2

Resblock from EDSR [72] is taken as baseline. FLOPs are calculated on an image of size 256×256. *'g' represents the number of groups in the group convolutions.*

TABLE 9
Feature Aggregation

	Sum	Concat	SKFF
PSNR (in dB)	30.76	30.83	30.97
Parameters	0	8,192	1,536

Our SKFF uses $\sim 5\times$ *fewer parameters than 'Concat', but generates better results.*

TABLE 10
Effect of Progressive Learning

Patch size	128	144	192	224	Progressive
PSNR (in dB)	30.97	30.99	31.02	31.08	31.06
Train time (h)	14	17	25	33	22

For progressive training, we gradually increase image patch size from 128×128 *to* 224×224.

EDSR [72]. Moreover, removing the transform part from RCB causes drop in accuracy. Table 8 also shows that replacing the group convolutions with regular convolutions in RCB increases the PSNR score, but at the cost of significant increase in parameters and FLOPs. Therefore, we opt for RCB with group convolutions (g=2) as a balanced choice.

Next, we analyze the feature aggregation strategy in Table 9. It shows that the proposed SKFF generates favorable results compared to summation and concatenation. Note that our proposed SKFF module uses $\sim 5\times$ fewer parameters than concatenation. Table 10 shows that the progressive learning strategy on mixed-size image patches yields PSNR similar to the model trained on large image patches (ps=224), but takes less time for training. Finally, in Table 11 we study how the number of convolutional streams and columns (RCB blocks) of MRB affect the image restoration quality. We note that increasing the number of streams provides significant improvements, thereby justifying the importance of multi-scale features processing. Moreover, increasing the number of columns yields better scores, thus indicating the significance of information exchange among parallel streams for feature consolidation.

TABLE 11
Ablation Study on Different Layouts of MRB

PSNR	Cols = 1	Cols = 2	Cols = 3
Rows = 1	30.01	30.29	30.47
Rows = 2	30.65	30.79	30.85
Rows = 3	30.73	30.97	31.03

Rows *denote the number of parallel resolution streams, and* Cols *represent the number of columns containing RCBs.*

5 Concluding Remarks

Conventional image restoration and enhancement pipelines either stick to the full resolution features along the network hierarchy or use an encoder-decoder architecture. The first approach helps retain precise spatial details, while the latter one provides better contextualized representations. However, these methods can satisfy only one of the above two requirements, although real-world image restoration tasks demand a combination of both conditioned on the given input sample. In this work, we propose a novel architecture whose main branch is dedicated to full-resolution processing and the complementary set of parallel branches provides better contextualized features. We propose novel mechanisms to learn relationships between features within each branch as well as across multi-scale branches. Our feature fusion strategy ensures that the receptive field can be dynamically adapted without sacrificing the original feature details. Consistent achievement of state-of-the-art results on six datasets for four image restoration and enhancement tasks corroborates the effectiveness of our approach.

References

[1] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervention*, 2015, pp. 234–241.

[2] O. Kupyn, T. Martyniuk, J. Wu, and Z. Wang, "DeblurGAN-v2: Deblurring (orders-of-magnitude) faster and better," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 8877–8886.

[3] C. Chen, Q. Chen, J. Xu, and V. Koltun, "Learning to see in the dark," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3291–3300.

[4] Y. Zhang, J. Zhang, and X. Guo, "Kindling the darkness: A practical low-light image enhancer," in *Proc. 27th ACM Int. Conf. Multimedia*, 2019, pp. 1632–1640.

[5] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2016.

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

[7] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image restoration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 7, pp. 2480–2495, Jul. 2021.

[8] A. Ignatov, N. Kobyshev, R. Timofte, K. Vanhoey, and L. Van Gool , "DSLR-quality photos on mobile devices with deep convolutional networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2017, pp. 3297–3305.

[9] S. W. Zamir *et al.*, "Learning enriched features for real image restoration and enhancement," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 492–511.

[10] A. Abdelhamed, S. Lin, and M. S. Brown, "A high-quality denoising dataset for smartphone cameras," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1692–1700.

[11] 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.*, 2018, pp. 8174–8182.

[12] S. Nah, T. H. Kim, and K. M. Lee, "Deep multi-scale convolutional neural network for dynamic scene deblurring," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 257–265.

[13] S. Gu, Y. Li, L. V. Gool, and R. Timofte, "Self-guided network for fast image denoising," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 2511–2520.

[14] A. Abuolaim and M. S. Brown, "Defocus deblurring using dual-pixel data," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 111126.

[15] 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.

[16] A. Abuolaim, M. Delbracio, D. Kelly, M. S. Brown, and P. Milanfar, "Learning to reduce defocus blur by realistically modeling dual-pixel data," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 2269–2278.

[17] 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. 2017.

[18] J. Shi, L. Xu, and J. Jia, "Just noticeable defocus blur detection and estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 657–665.

[19] J. Lee, S. Lee, S. Cho, and S. Lee, "Deep defocus map estimation using domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 12214–12222.

[20] L. P. Yaroslavsky, "Local adaptive image restoration and enhancement with the use of DFT and DCT in a running window," in *Wavelet Applications Signal Image Processing IV*, Bellingham, WA, USA: SPIE, 1996.

[21] E. P. Simoncelli and E. H. Adelson, "Noise removal via Bayesian wavelet coring," in *Proc. IEEE Int. Conf. Image Process.*, 1996, pp. 379–382.

[22] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 1998, pp. 839–846.

[23] P. Perona and J. Malik, "Scale-space and edge detection using anisotropic diffusion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 12, no. 7, pp. 629–639, Jul. 1990.

[24] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Phys. D: Nonlinear Phenomena*, vol. 60, pp. 259–268, 1992.

[25] A. A. Efros and T. K. Leung, "Texture synthesis by non-parametric sampling," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 1999, pp. 1033–1038.

[26] A. Buades, B. Coll, and J.-M. Morel, "A non-local algorithm for image denoising," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2005, pp. 60–65.

[27] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image denoising by sparse 3-D transform-domain collaborative filtering," *IEEE Trans. Image Process.*, vol. 16, no. 8, pp. 2080–2095, Aug. 2007.

[28] W. Dong, G. Shi, and X. Li, "Nonlocal image restoration with bilateral variance estimation: A low-rank approach," *IEEE Trans. Image Process.*, vol. 22, no. 2, pp. 700–711, Feb. 2013.

[29] S. Gu, L. Zhang, W. Zuo, and X. Feng, "Weighted nuclear norm minimization with application to image denoising," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 2862–2869.

[30] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Non-local sparse models for image restoration," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2009, pp. 2272–2279.

[31] R. Hedjam, R. F. Moghaddam, and M. Cheriet, "Markovian clustering for the non-local means image denoising," in *Proc. IEEE Int. Conf. Image Process.*, 2009, pp. 3877–3880.

[32] S. Anwar and N. Barnes, "Real image denoising with feature attention," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 3155–3164.

[33] T. Brooks, B. Mildenhall, T. Xue, J. Chen, D. Sharlet, and J. T. Barron, "Unprocessing images for learned raw denoising," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 11028–11037.

[34] S. Guo, Z. Yan, K. Zhang, W. Zuo, and L. Zhang, "Toward convolutional blind denoising of real photographs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1712–1722.

[35] T. Plötz and S. Roth, "Neural nearest neighbors networks," in *Proc. 32nd Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 1095–1106.

[36] K. Zhang, W. Zuo, and L. Zhang, "FFDNet: Toward a fast and flexible solution for CNN-Based image denoising," *IEEE Trans. Image Process.*, vol. 27, no. 9, pp. 4608–4622, Sep. 2018.

[37] S. W. Zamir *et al.*, "CycleISP: Real image restoration via improved data synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 2693–2702.

[38] M. Chang, Q. Li, H. Feng, and Z. Xu, "Spatial-adaptive network for single image denoising," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 171–187.

[39] Z. Yue, Q. Zhao, L. Zhang, and D. Meng, "Dual adversarial network: Toward real-world noise removal and noise generation," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 41–58.

[40] Z. Yue, H. Yong, Q. Zhao, D. Meng, and L. Zhang, "Variational denoising network: Toward blind noise modeling and removal," in *Proc. 32nd Int. Conf. Neural Inf. Process. Syst.*, 2019, pp. 1690–1701.

[41] Y. Kim, J. W. Soh, G. Y. Park, and Nam I. Cho, "Transfer learning from synthetic to real-noise denoising with adaptive instance normalization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3479–3489.

[42] F. Fang, J. Li, Y. Yuan, T. Zeng, and G. Zhang, "Multilevel edge features guided network for image denoising," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 9, pp. 3956–3970, Sep. 2021.

[43] R. Keys, "Cubic convolution interpolation for digital image processing," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 29, no. 6, pp. 1153–1160, Dec. 1981.

[44] M. Irani and S. Peleg, "Improving resolution by image registration," *CVGIP: Graphical Models Image Process.*, vol. 53, pp. 231–239, 1991.

[45] J. Allebach and P. W. Wong, "Edge-directed interpolation," in *Proc. IEEE Int. Conf. Image Process.*, 1996, pp. 707–710.

[46] L. Zhang and X. Wu, "An edge-guided image interpolation algorithm via directional filtering and data fusion," *IEEE Trans. Image Process.*, vol. 15, no. 8, pp. 2226–2238, Aug. 2006.

[47] K. I. Kim and Y. Kwon, "Single-image super-resolution using sparse regression and natural image prior," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 6, pp. 1127–1133, Jun. 2010.

[48] Z. Xiong, X. Sun, and F. Wu, "Robust web image/video super-resolution," *IEEE Trans. Image Process.*, vol. 19, no. 8, pp. 2017–2028, Aug. 2010.

[49] H. Chang, D.-Y. Yeung, and Y. Xiong, "Super-resolution through neighbor embedding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2004, pp. I–I.

[50] G. Freedman and R. Fattal, "Image and video upscaling from local self-examples," *ACM Trans. Graph.*, vol. 20, 2011, Art. no. 12.

[51] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Trans. Image Process.*, vol. 19, no. 11, pp. 2861–2873, Nov. 2010.

[52] J. Yang, J. Wright, T. Huang, and Y. Ma, "Image super-resolution as sparse representation of raw image patches," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2008, pp. 1–8.

[53] Z. Wang, J. Chen, and S. C. H. Hoi, "Deep learning for image super-resolution: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3365–3387, Oct. 2021.

[54] S. Anwar, S. Khan, and N. Barnes, "A deep journey into super-resolution: A survey," 2019, *arXiv*.

[55] J. Cai, S. Gu, R. Timofte, and L. Zhang, "Ntire 2019 challenge on real image super-resolution: Methods and results," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2019, pp. 2211–2223.

[56] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 184–199.

[57] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2016, pp. 1646–1654.

[58] Y. Tai, J. Yang, X. Liu, and C. Xu, "MemNet: A persistent memory network for image restoration," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2017, pp. 4549–4557.

[59] Y. Tai, J. Yang, and X. Liu, "Image super-resolution via deep recursive residual network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2790–2798.

[60] Z. Hui, X. Wang, and X. Gao, "Fast and accurate single image super-resolution via information distillation network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 723–731.

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

[62] J. Kim, J. K. Lee, and K. Mu Lee , "Deeply-recursive convolutional network for image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1637–1645.

[63] W. Han, S. Chang, D. Liu, M. Yu, M. Witbrock, and T. S. Huang, "Image super-resolution via dual-state recurrent networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1654–1663.

[64] N. Ahn, B. Kang, and K.-A. Sohn, "Fast, accurate, and lightweight super-resolution with cascading residual network," in *Proc. Eur. Conf. Comput. Vis.*, 2018.

[65] Z. Wang, D. Liu, J. Yang, W. Han, and T. Huang, "Deep networks for image super-resolution with sparse prior," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2015, pp. 370–378.

[66] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Deep laplacian pyramid networks for fast and accurate superresolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5835–5843.

[67] T. Tong, G. Li, X. Liu, and Q. Gao, "Image super-resolution using dense skip connections," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2017, pp. 4809–4817.

[68] X. Wang *et al.*, "ESRGAN: Enhanced super-resolution generative adversarial networks," in *Proc. Eur. Conf. Comput. Vis. Workshops*, 2018, pp. 63–79.

[69] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 219–224.

[70] T. Dai, J. Cai, Y. Zhang, S.-T. Xia, and L. Zhang, "Second-order attention network for single image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 11057–11066.

[71] Y. Zhang, K. Li, K. Li, B. Zhong, and Y. Fu, "Residual non-local attention networks for image restoration," in *Proc. Int. Conf. Learn. Representations*, 2019.

[72] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee , "Enhanced deep residual networks for single image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2017, pp. 1132–1140.

[73] R. Dahl, M. Norouzi, and J. Shlens, "Pixel recursive super resolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2017, pp. 5449–5458.

[74] J. Li, F. Fang, K. Mei, and G. Zhang, "Multi-scale residual network for image super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 2006–2013.

[75] S.-J. Park, H. Son, S. Cho, K.-S. Hong, and S. Lee, "SRFEAT: Single image super-resolution with feature discrimination," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 455–471.

[76] M. S. M. Sajjadi, B. Scholkopf, and M. Hirsch, "Enhancenet: Single image super-resolution through automated texture synthesis," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2017, pp. 4501–4510.

[77] C. Ledig *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 105–114.

[78] E. H. Land, "The retinex theory of color vision," *Sci. Amer.*, vol. 237, pp. 108–129, 1977.

[79] M. Bertalmío, V. Caselles, E. Provenzi, and A. Rizzi, "Perceptual color correction through variational techniques," *IEEE Trans. Image Process.*, vol. 16, no. 4, pp. 1058–1072, Apr. 2007.

[80] R. Palma-Amestoy , E. Provenzi, M. Bertalmío, and V. Caselles, "A perceptually inspired variational framework for color enhancement," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 3, pp. 458–474, Mar. 2009.

[81] D. J. Jobson, Z.-U. Rahman, and G. A. Woodell, "A multiscale retinex for bridging the gap between color images and the human observation of scenes," *IEEE Trans. Image Process.*, vol. 6, no. 7, pp. 965–976, Jul. 1997.

[82] A. Rizzi, C. Gatta, and D. Marini, "From retinex to automatic color equalization: Issues in developing a new algorithm for unsupervised color equalization," in *Journal Electronic Imaging*, Bellingham, WA, USA: SPIE, 2004.

[83] A. Ignatov and R. Timofte, "NTIRE 2019 challenge on image enhancement: Methods and results," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2019, pp. 2224–2232.

[84] L. Shen, Z. Yue, F. Feng, Q. Chen, S. Liu, and J. Ma, "MSR-net: Low-light image enhancement using deep convolutional network," *arXiv*, 2017.

[85] C. Wei, W. Wang, W. Yang, and J. Liu, "Deep retinex decomposition for low-light enhancement," in *Proc. Brit. Mach. Vis. Conf.*, 2018.

[86] H. Chang, M. K. Ng, W. Wang, and T. Zeng, "Retinex image enhancement via a learned dictionary," *Optical Engineering*, Bellingham, WA, USA: SPIE, 2015.

[87] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 833–851.

[88] K. G. Lore, A. Akintayo, and S. Sarkar, "LLNet: A deep autoencoder approach to natural low-light image enhancement," *Pattern Recognit.*, vol. 61, pp. 650–662, 2017.

[89] W. Ren *et al.*, "Low-light image enhancement via a deep hybrid network," *IEEE Trans. Image Process.*, vol. 29, no. 9, pp. 4364–4375, Sep. 2019.

[90] K. Mei, J. Li, J. Zhang, H. Wu, J. Li, and R. Huang, "Higher-resolution network for image demosaicing and enhancing," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop*, 2019, pp. 3441–3448.

[91] J. Li, J. Li, F. Fang, F. Li, and G. Zhang, "Luminance-aware pyramid network for low-light image enhancement," *IEEE Trans. Multimedia*, vol. 23, pp. 3153–3165, 2020.

[92] Y.-S. Chen, Y.-C. Wang, M.-H. Kao, and Y.-Y. Chuang, "Deep photo enhancer: Unpaired learning for image enhancement from photographs with gans," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6306–6314.

[93] A. Ignatov, N. Kobyshev, R. Timofte, K. Vanhoey, and L. Van Gool , "WESPE: Weakly supervised photo enhancer for digital cameras," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2018, pp. 804–809.

[94] Y. Deng, C. C. Loy, and X. Tang, "Aesthetic-driven image enhancement by adversarial learning," in *Proc. 26th ACM Int. Conf. Multimedia*, 2018, pp. 870–878.

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

[96] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 483–499.

[97] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2015, pp. 1520–1528.

[98] B. Xiao, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 472–487.

[99] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[100] X. Peng, R. S. Feris, X. Wang, and D. N. Metaxas, "A recurrent encoder-decoder network for sequential face alignment," in *Proc. Eur. Conf. Comput. Vis.*, 2016.

[101] D. H. Hubel and T. N. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," *J. Physiol.*, vol. 160, pp. 106154, 1962.

[102] M. Riesenhuber and T. Poggio, "Hierarchical models of object recognition in cortex," *Nature Neurosci.*, vol. 2, pp. 1019–1025, 1999.

[103] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio, "Robust object recognition with cortex-like mechanisms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 3, pp. 411–426, Mar. 2007.

[104] C. P. Hung, G. Kreiman, T. Poggio, and J. J. DiCarlo , "Fast read-out of object identity from macaque inferior temporal cortex," *Science*, vol. 310, pp. 863–866, 2005.

[105] G. Huang, D. Chen, T. Li, F. Wu, L. van der Maaten, and K. Q. Weinberger, "Multi-scale dense networks for resource efficient image classification," in *Proc. Int. Conf. Learn. Representations*, 2018.

[106] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5686–5696.

[107] D. Fourure, R. Emonet, É. Fromont, D. Muselet, A. Trémeau, and C. Wolf, "Residual conv-deconv grid network for semantic segmentation," in *Proc. Brit. Mach. Vis. Conf.*, 2017.

[108] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.

[109] X. Li, W. Wang, X. Hu, and J. Yang, “Selective kernel networks,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 510–519.

[110] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[111] X. Wang, R. Girshick, A. Gupta, and K. He, “Non-local neural networks,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7794–7803.

[112] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, “Global context networks,” *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Dec. 24, 2020, doi: 10.1109/TPAMI.2020.3047209.

[113] E. Hoffer, B. Weinstein, I. Hubara, T. Ben-Nun , T. Hoefler, and D. Soudry, “Mix & match: Training convnets with mixed image sizes for improved accuracy, speed and scale resiliency,” 2019, *arXiv:1908.08986*.

[114] M. Tan and Q. Le, “Efficientnet: Rethinking model scaling for convolutional neural networks,” in *Proc. Int. Conf. Mach. Learn.*, 2019.

[115] T. Plotz and S. Roth, “Benchmarking denoising algorithms with real photographs,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2750–2759.

[116] 2017, Accessed: Feb. 29, 2020. [Online]. Available: https://noise.visinf.tu-darmstadt.de/benchmark/

[117] J. Cai, H. Zeng, H. Yong, Z. Cao, and L. Zhang, “Toward real-world single image super-resolution: A new benchmark and a new model,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 3086–3095.

[118] V. Bychkovsky, S. Paris, E. Chan, and F. Durand, “Learning photographic global tonal adjustment with a database of input/output image pairs,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 97–104.

[119] Y. Hu, H. He, C. Xu, B. Wang, and S. Lin, “Exposure: A white-box photo post-processing framework,” *ACM Trans. Graph.*, vol. 37, 2018.

[120] J. Park, J.-Y. Lee, D. Yoo, and I. So Kweon , “Distort-and-recover: Color enhancement using deep reinforcement learning,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 5928–5936.

[121] R. Wang, Q. Zhang, C.-W. Fu, X. Shen, W.-S. Zheng, and J. Jia, “Underexposed photo enhancement using deep illumination estimation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 6842–6850.

[122] I. Loshchilov and F. Hutter, “SGDR: Stochastic gradient descent with warm restarts,” in *Proc. Int. Conf. Learn. Representations*, 2017.

[123] H. C. Burger, C. J. Schuler, and S. Harmeling, “Image denoising: Can plain neural networks compete with BM3D?,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 2392–2399.

[124] C. Mou, J. Zhang, and Z. Wu, “Dynamic attentive graph learning for image restoration,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 4308–4317.

[125] C. Ren, X. He, C. Wang, and Z. Zhao, “Adaptive consistency prior based deep network for image denoising,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 8592–8602.

[126] Z. Ying, G. Li, and W. Gao, “A bio-inspired multi-exposure fusion framework for low-light image enhancement,” 2017, *arXiv:1711.00591*.

[127] Z. Ying, G. Li, Y. Ren, R. Wang, and W. Wang, “A new image contrast enhancement algorithm using exposure fusion framework,” in *Proc. Int. Conf. Comput. Anal. Images Patterns*, 2017.

[128] X. Dong, G. Wang, Y. Pang, W. Li, J. Wen, W. Meng, and Y. Lu, “Fast efficient algorithm for enhancement of low lighting video,” in *Proc. IEEE Int. Conf. Multimedia Expo*, 2011, pp. 1–6.

[129] X. Guo, Y. Li, and H. Ling, “LIME: Low-light image enhancement via illumination map estimation,” *IEEE Trans. Image Process.*, vol. 26, no. 2, pp. 982–993, Feb. 2017.

[130] X. Fu, D. Zeng, Y. Huang, X.-P. Zhang, and X. Ding, “A weighted variational model for simultaneous reflectance and illumination estimation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2782–2790.

[131] Y. Liu, R. Wang, S. Shan, and X. Chen, “Structure inference net: Object detection using scene-level context and instance-level relationships,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6985–6994.

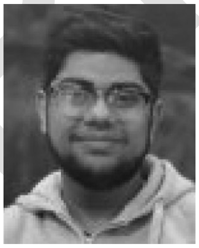
[132] S. Wang, J. Zheng, H.-M. Hu, and B. Li, “Naturalness preserved enhancement algorithm for non-uniform illumination images,” *IEEE Trans. Image Process.*, vol. 22, no. 9, pp. 3538–3548, Sep. 2013.

[133] W. Wang, C. Wei, W. Yang, and J. Liu, “GLADNet: Low-light enhancement network with global awareness,” in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2018, pp. 751–755.

[134] Y. Zhang, X. Guo, J. Ma, W. Liu, and J. Zhang, “Beyond brightening low-light images,” *Int. J. Comput. Vis.*, vol. 129, pp. 1013–1037, 2021.

[135] M. Gharbi, J. Chen, J. T. Barron, S. W. Hasinoff, and F. Durand, “Deep bilateral learning for real-time image enhancement,” *ACM Trans. Graph.*, vol. 118, 2017, Art. no. 118.

Syed Waqas Zamir received the PhD degree from the University Pompeu Fabra, Barcelona, Spain, in 2017. He is a research scientist with the Inception Institute of Artificial Intelligence, UAE. His research interests include low-level computer vision, computational imaging, image and video processing, color vision and image restoration and enhancement.



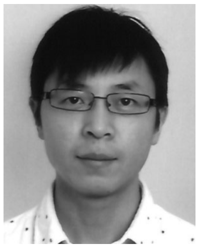
Aditya Arora is a research engineer with the Inception Institute of Artificial Intelligence, UAE. His research interests include image and video processing, computational photography, and low-level vision.

Salman Khan received the PhD degree from the University of Western Australia, Perth, Australia, in 2016. He is an assistant professor with the MBZ University of Artificial Intelligence. He has been an adjunct faculty member with Australian National University since 2016. He has been awarded the outstanding reviewer award at CVPR multiple times, won the Best Paper Award at 9th ICPRAM 2020, and 2nd prize in the NTIRE Image Enhancement Competition at CVPR 2019. He served as a program committee member for several premier conferences including CVPR, ICCV, ICLR, ECCV, and NeurIPS. His thesis received an honorable mention on the Dean’s List Award. His research interests include computer vision and machine learning.

Munawar Hayat received the PhD degree from the University of Western Australia (UWA), Perth, Australia. His PhD thesis received multiple awards, including the Deans List Honorable Mention Award and the Robert Street Prize. After his PhD, he joined IBM Research as a postdoc and then moved to the University of Canberra as an assistant professor. He is currently a senior scientist with the Inception Institute of Artificial Intelligence, UAE. He was granted two U.S. patents, and has published more than 30 papers at leading venues in his field, including the *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *International Journal of Computer Vision*, CVPR, ECCV, and ICCV. His research interests include computer vision and machine/deep learning.

Fahad Shahbaz Khan received the MSc degree in intelligent systems design from the Chalmers University of Technology, Gothenburg, Sweden, and the PhD degree in computer vision from the Autonomous University of Barcelona, Bellaterra, Spain. He is a faculty member with MBZUAI, UAE and Linkoping University, Sweden. From 2018 to 2020, he worked as a lead scientist with the Inception Institute of Artificial Intelligence (IIAI), Abu Dhabi, United Arab Emirates. He has achieved top ranks on various international challenges (Visual Object Tracking VOT: 1st 2014 and 2018, 2nd 2015, 1st 2016; VOT-TIR: 1st 2015 and 2016; OpenCV Tracking: 1st 2015; 1st PASCAL VOC 2010). His research interests include a wide range of topics within computer vision and machine learning, such as object recognition, object detection, action recognition, and visual tracking. He has published articles in high-impact computer vision journals and conferences in these areas. He serves as a regular program committee member for leading computer vision conferences such as CVPR, ICCV, and ECCV.

Ming-Hsuan Yang (Fellow, IEEE) is affiliated with Google, UC Merced, and Yonsei University. He serves as a program co-chair of IEEE International Conference on Computer Vision (ICCV) in 2019, program co-chair of Asian Conference on Computer Vision (ACCV) in 2014, and general co-chair of ACCV 2016. He served as an associate editor of the *IEEE Transactions on Pattern Analysis and Machine Intelligence*, and is an associate editor of the *International Journal of Computer Vision*, *Image and Vision Computing* and *Journal of Artificial Intelligence Research*. He received the NSF CAREER award and Google Faculty Award.



Ling Shao (Fellow, IEEE) is the chief scientist with Terminus Group and the president of Terminus International. He was the founding CEO and chief scientist with the Inception Institute of Artificial Intelligence, Abu Dhabi, UAE. His research interests include computer vision, deep learning, medical imaging and vision and language. He is a fellow of the IAPR, BCS, and IET.

▷ **For more information on this or any other computing topic, please visit our Digital Library at** www.computer.org/csdl.