


ORIGINAL RESEARCH PAPER

Image quality enhancement using hybrid attention networks

Jiachen Wang  | Yingyun Yang | Yan Hua

State Key Laboratory of Media Convergence and
Communication, Communication University of
China, Beijing, China

Correspondence

Yingyun Yang, State Key Laboratory of Media
Convergence and Communication, Communication
University of China, No. 1 Dingfuzhuang East Street,
Chaoyang District, Beijing, China
Email: yangyingyun@cuc.edu.cn

Funding information

National Key R&D Program of China, Grant/Award
Number: 2020YFB1406800

Abstract

Image quality enhancement aims to recover rich details from degraded images, which is applied into many fields, such as medical imaging, filming production and autonomous driving. Deep convolutional neural networks (CNNs) have enabled rapid development of image quality enhancement. However, most existing CNN-based methods lack versatility when targeting different subtasks in terms of the design of networks. Besides, they often fail to balance precise spatial representations and necessary contextual information. To deal with these problems, this paper proposes a novel unified framework for low-light image enhancement, image denoising and image super-resolution. The core of this architecture is a residual hybrid attention block (RHAB), which consists of several dynamic down-sampling modules (DDM) and hybrid attention up-sampling modules (HAUM). Specifically, multi-scale feature maps are fully interacted with each other with the help of nested subnetworks so that both high-resolution spatial details and high-level contextual information can be combined to improve the representation ability of the network. Further, a hybrid attention network (HAN) is proposed and evaluations on three separate subtasks demonstrate its good performance. Extensive experiments on the authors' synthetic dataset, a more complex version, show that the authors' method achieve better quantitative and visual results compared to other state-of-the-art methods.

1 | INTRODUCTION

Images with high quality are essential to people's daily life and production. However, an image is usually very susceptible to surrounding environments during its acquisition, transmission, and processing. For example, poor light conditions, noise caused by devices and compression damage during transmission may make the image suffer from serious information loss and reduce its quality. Low-quality images not only fail to meet the visual requirements of human eyes, but also affect the subsequent processing and analysis by computers. Therefore, there is an urgent need to recover the necessary content and details from degraded images.

Recently, great advances [1–14] have been made to improve the quality of images and most of them are based on deep convolutional neural networks (CNNs). This paper defines image quality enhancement as low-light image enhancement, image super-resolution and image denoising. Although a lot of researches have been invested into exploring more novel

and robust network architecture. However, there still exist two problems.

First, the design ideas of networks for different subtasks are quite different. For example, retinex theory [15], which assumes that an image can be regarded as the product of reflection and illumination, plays a vital role in low-light image enhancement techniques. Recent algorithms [16–20] combining retinex theory and deep CNNs try to enhance the reflection instead of the image itself and show outstanding restoration and enhancement ability compared to other approaches. However, recent image denoising methods [21–24] regard a noisy image as the sum of a clean image and the noise. They aim to learn the residual between noisy images and the noise. In the field of image super-resolution, mainstream approaches [25–29] proposed in recent years first learn the low-resolution representations with deep CNNs and then reconstruct the high-resolution image by the operation of pixel shuffle [30]. Therefore, the network designed for a certain independent subtask is difficult to be directly applied to other tasks. In addition, real-world

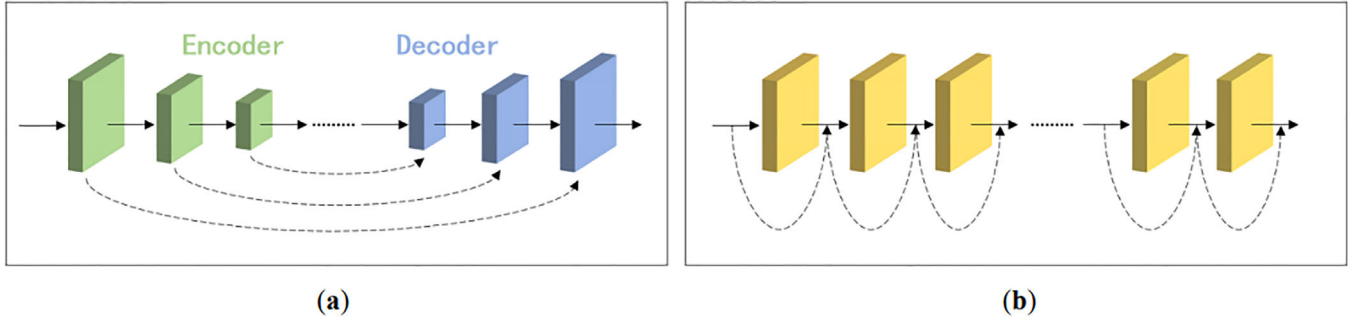


FIGURE 1 Two main categories of architecture designs. (a) Encoder-decoder structure; (b) Single-resolution structure

low-quality images are often corrupted by multiple degradation kernels, which means they could be low-light, low-resolution and noisy simultaneously. Existing methods can hardly be used to handle such a complex task.

Besides, existing CNN-based methods often fail to balance precise spatial representations and necessary contextual information. Existing deep CNNs for image quality enhancement can be divided into two categories in terms of their architecture designs. One is encoder-decoder structure (Figure 1a), represented by U-net [31] and fully convolution networks (FCNs) [32], and the other is single-resolution structure (Figure 1b). The former one passes the input to an encoder to learn low-resolution representations and then high-resolution output is reconstructed with a decoder. Encoder-decoder networks [17, 19, 22, 33-37] are usually capable of increasing the receptive field so as to capture richer contextual information, which is semantic and can improve the reconstruction to some extent. Nevertheless, it is inevitable that some details are not precise because of the low-to-high resolution reconstruction process. And this may cause spatial inaccuracy in restored images. Single-resolution models [3, 23, 38-40] keep the resolution of feature maps unchanged during the whole process. This strategy is applicable when the size of input image is small. However, if the fixed resolution is too high, large-sized feature maps must be maintained, which may take up a lot of memory resources and slow down the inference speed.

In order to solve the problem that networks lack versatility in different tasks, we do not follow any task-specific pipeline and propose an end-to-end trainable model called hybrid attention network (HAN). Given an input low-quality image, it will directly generate the enhanced result without estimating the reflectance, illumination, or noise. This means that our network must have stronger representation abilities to make up for the performance degradation caused by lack of the module designed for specific task. To achieve this goal, our HAN is composed of several strong residual hybrid attention blocks (RHABs). Each RHAB is built on hybrid attention up-sampling modules (HAUMs), dynamic down-sampling modules (DDMs) and nested pathways. It maintains three multi-resolution branches, where cross-scale features could fully interact with each other with the help of

HAUMs and DDMs. Therefore, the design of RHAB makes it possible to efficiently extract contextual information without losing fine-grained details so that shortcomings of encoder-decoder structure and single-resolution structure are tackled. RHAB is a strong module and also a key innovation in our method.

Experiments are conducted on several benchmark datasets for separate subtasks. Quantitative and qualitative results show that our method can achieve the best or the second-best performance in most subtasks. Further, we synthesize a more complex dataset, in which images are extremely corrupted to low-light, low-resolution and noisy versions. Experiments on this tough task demonstrate that our HAN achieves best performance both quantitatively and visually compared to other state-of-the-art methods.

In summary, the main contributions of our work are as follows:

1. We propose a novel residual hybrid attention block (RHAB) and further establish a hybrid attention network (HAN) for image quality enhancement.
2. Our proposed network can achieve good performance in independent subtasks and outperform other state-of-the-art methods in the face of multiple degradation kernels.

The rest of our paper is organized as follows. In Section 2, both related methods in the field of image quality enhancement and efficient strategies that inspire us are briefly introduced. In Section 3, we present our proposed method in details. Experimental results and ablation study are shown in Sections 4 and 5, respectively. We also discuss the difference between our network and others in Section 6. Finally, in Section 7, conclusions are drawn and future work is given.

2 | RELATED WORK

Due to the rapid development of deep learning techniques, CNN-based methods in image-quality-enhancement related fields have become the mainstream in recent years. Therefore, in this section, deep CNN-based algorithms are mainly introduced.

2.1 | Image quality enhancement

This paper defines image quality enhancement as a comprehensive task including low-light image enhancement, image denoising and image super-resolution.

Low-light images are usually produced by poor lighting conditions or inappropriate camera parameters. They lack enough contrast and it is difficult to recognize what they contain. CNNs have achieved great success in low-light image enhancement since Lore et al. [33] proposed a CNN-based method in this field. This network is based on a stacked sparse autoencoder. Tao et al. refer to ResNet [41] and GoogleNet [42], and propose LLCNN [38], where all convolutional layers operate on single resolution. Motivated by retinex theory [15] and multi-scale retinex (MSR) [43], Shen et al. [16] employ convolutional layers to imitate the MSR process and what they propose achieves great performance. Wei et al. [17] also introduce retinex theory into the network design. They adopt a multi-step strategy, that is, decomposing the low-light image into reflection and illumination, enhancing the reflectance, and reconstructing the normal-light image. Their method is further improved by Zhang et al. [18] but retinex-based network architecture is still preserved. Chen et al. [34] propose an end-to-end network based on fully convolution networks (FCNs) to enhance images in RAW space and achieve outstanding success. Jiang et al. [20] propose an attention guided U-net to generate enhanced image, but it is also based on retinex theory. The retinex theory still plays a vital role in recent researches [44, 45].

The noise in images is generated in various processes of capturing, transmission, compressing and processing. Early trials [46, 47] of introducing deep learning into image denoising fail to achieve better performance compared to traditional methods because their networks are relatively too simple. DnCNN [23], a CNN-based denoiser, obtains significant performance gains and successfully proves the advantages of deep CNNs. It is a single-resolution network with batch normalization (BN) [48] and residual structure. In order to enlarge the receptive field of single-resolution network, Anwar et al. [21] introduce dilation convolution [49] into their method. Guo et al. propose CBDNet [22] consisting of a noise estimation subnetwork and a denoising subnetwork. The former is a single-resolution structure while the latter is an encoder-decoder structure. Similar network design is also proposed in VDN [50]. Nearly all methods regard a noisy image as the sum of a clean image and the noise.

Low-resolution images are also generated during the period of image processing. The core of image super-resolution lies in how to reconstruct lost details. SRCNN [51] is the first CNN proposed for this task. It is a single resolution network with only three layers. Although some methods [35, 52] are based on encoder-decoder structure, single-resolution networks have been becoming the mainstream since sub-pixel convolution [30] was proved to be a more efficient way of up-sampling. Among them, VDSR [53], consisting of twenty layers, predicts the residual between the up-sampled low-resolution input and high-resolution ground truth. EDSR [26] achieves great success by removing BN layers and increasing both the width and

depth of the network. SRDensenet [28] introduces dense skip connections [54] into image super-resolution for feature reuse. RDN [3], combining residual learning and dense skip connections, achieves the state-of-the-art performance. It is also a single-resolution network. Recent image super-resolution methods [55, 56] also follow the thought that features are first extracted using single-resolution networks and upsampled in the end. Different from the tasks of low-light image enhancement and image denoising, a low-resolution image is degraded by more complex kernels that cannot be easily formulated with addition or multiplication. Hence, most researches try to improve the performance by improving the non-linear ability of networks. Increasing the depth of networks proves to be a useful way.

It can be concluded that the reconstruction processes in each task are quite different. For example, it is obviously unreasonable to apply the retinex-based methods into image super-resolution. So far, there has been very few researches dedicated to a unified framework that can handle all subtasks in image quality enhancement. Also, an image may be corrupted with multiple degradations and there is an urgent need to propose a network that can enhance low-light, noisy and low-resolution images in one step. Besides, recent CNN-based methods are mainly built on encoder-decoder structure and single-resolution structure. It is not easy for encoder-decoder networks to ensure spatial accuracy during the process of up-sampling, while single-resolution networks are not efficient enough under the circumstance when the resolution of input is too high. Therefore, we try to propose a novel network that can apply to multiple tasks. It should also have the advantages of both encoder-decoder structure and single-resolution structure.

2.2 | Multi-resolution strategy

Some researchers make efforts on a trade-off between encoder-decoder structure and single-resolution structure. They try to propose a multi-resolution structure that can overcome the disadvantages of above two structures. Recently, Sun et al. propose the HRNet [57] for human pose estimation. In HRNet, both high-resolution representations and low-resolution representations are maintained. Multi-resolution features are fused and separated every a few layers. The down-sampling in HRNet is implemented by strided convolution while the up-sampling is implemented by nearest neighbour interpolation and 1×1 convolution. Despite HRNet maintains parallel multi-resolution branches, the way of cross-scale feature fusion is still very simple and may cause distortion to some extent. Some improvements are also made to give full play to HRNet. For example, Huo et al. [58] adopt the attention to suppress the unimportant information and Wang et al. [59] additionally consider the multi-level semantic information. In the task of image restoration and enhancement, Zamir et al. follow such design principle and further propose the MIRNet [2]. In terms of its way of down-sampling, anti-aliasing down-sampling [60] is adopted. Qin [61] et al. also propose a similar network based on multi-resolution strategy while the attention is considered in the process of

feature fusion. We follow this multi-resolution strategy and propose a basic block containing multi-scale branches so that it can efficiently balance spatial accuracy and contextual information.

2.3 | Attention

SENet, proposed by Hu et al. [62] is a light-weight plug-in to calculate channel-wise attention. Zhang et al. adopt a similar channel attention mechanism and propose RCAN [29]. In CBAM [63], spatial attention is also taken into consideration subsequently after the channel attention. Further, Tian et al. [64] propose CANet for image restoration based on such concatenated attention module. Different from CBAM, dual attention [65] generates channel attention and spatial attention in parallel. The idea of dual attention is also introduced into MIRNet and plays a vital role in learning enriched features. In our method, a stronger attention mechanism is proposed. In the process of up-sampling, both channel attention and spatial attention are used to guide the interaction between multi-resolution feature maps, and to promote the fusion of different information.

3 | MATERIALS AND METHODS

In this section, detailed descriptions about our proposed Hybrid Attention Network (HAN) are given. The basic component of our HAN is the residual hybrid attention block (RHAB). In each RHAB, dynamic down-sampling modules (DDMs) are used to relieve information loss. To address the problem of inaccurate up-sampling, hybrid attention up-sampling modules (HAUMs) are also proposed. They work together to make the multi-scale feature interaction better.

3.1 | Overall pipeline

As shown in Figure 2, a low-quality image and its ground truth (a high-quality version) are denoted as I_{LQ} and I_{HQ} , respectively. Our goal is to generate a clean and high-resolution image with moderate illumination that is suitable for human perception or further processing by computers. Let's denote I_{EQ} as the output of the network. This reconstructed image should be as close as possible to I_{HQ} . Therefore, the pipeline of our method is given as follows:

$$I_{EQ} = \text{HAN} (I_{LQ}) \quad (1)$$

$$\min_{\theta} \mathcal{L} (I_{EQ}, I_{HQ}) \quad (2)$$

where $\text{HAN}(\cdot)$ and θ denote the operation and parameters of our network, respectively. $\mathcal{L}(\cdot)$ is the loss function used to measure the distance between two images. In order to avoid the problem of over-smoothing [66], we employ $\mathcal{L}1$ loss function in our network. So, given a training set with N image pairs, Equations (1) and (2) can be further formulated as:

tions (1) and (2) can be further formulated as:

$$\min_{\theta} \frac{1}{N} \sum_{i=1}^N \left\| \text{HAN} (I_{LQ}^i) - I_{HQ}^i \right\|_1 \quad (3)$$

where I_{LQ}^i and I_{HQ}^i denote the i -th low-quality image and its corresponding ground truth, respectively.

3.2 | Network architecture

Our HAN mainly consists of a series of residual hybrid attention blocks (RHABs). Feature maps extracted from these blocks are bypassed through short skip connections.

A long skip connection is also applied between the first and the last convolutional layer. This strategy has been proved effective in RCAN, as it can promote the hierarchical information flow.

To be specific, a degraded image I_{EQ} is first up-sampled with a scale factor f so that it will have the same size as desired output. In tasks where super-resolution is not needed, this process can be omitted. Like methods [3, 26], one convolutional layer is used to extract the shallow feature F_0 :

$$F_0 = C_S (U (I_{LQ})) \quad (4)$$

where $U(\cdot)$ and $C_S(\cdot)$ denote up-sampling and the operation of the first convolutional layer, respectively. Supposing C_D is the convolutional layer after the last RHAB, reconstructed feature F_R can be calculated with the following equation:

$$F_R = F_0 + C_D (F_b) \quad (5)$$

where F_b is the output of the last RHAB. Further, we utilize another convolutional layer C_G to generate I_{EQ} from F_R :

$$I_{EQ} = C_G (F_R) \quad (6)$$

where $C_G(\cdot)$ is the operation of the last convolutional layer in HAN.

RHAB is the key component of our HAN. It is a subnetwork based on residual learning [41] and has advantages of both encoder-decoder structure and single-resolution structure. As shown in Figure 2, F_{j-1} and F_j are respectively the input and output of the j -th RHAB. To simplify the learning progress, the purpose of each RHAB is to learn a residual between F_{j-1} and F_j , so a long skip connection within the block is established.

Our RHAB contains a high-resolution trunk and two low-resolution branches with $1/2$ and $1/4$ scale factors, respectively. In high-resolution trunk road, spatially accurate features are extracted while in low-resolution branches, rich global information is captured as a supplement of high-resolution features.

In order to tackle the issue of semantic gap [67] that may hinder the fusion of features with the same size, we introduce the nested pathways into RHAB. Besides, to extract

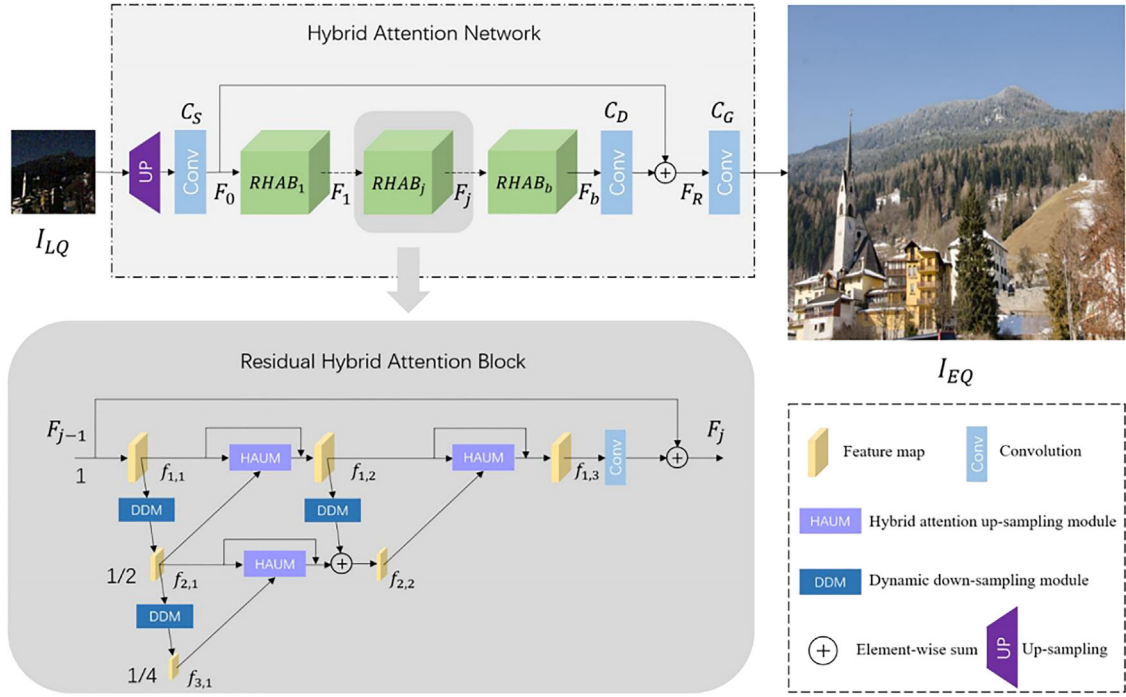


FIGURE 2 The overall architecture of hybrid attention network (HAN). We use bilinear interpolation as the up-sampling module before the first convolutional layer. Note that in some tasks where the input and output have the same size, this module is removed. Note that $f_{i,j}$ denotes the j -th feature map in i -th branch

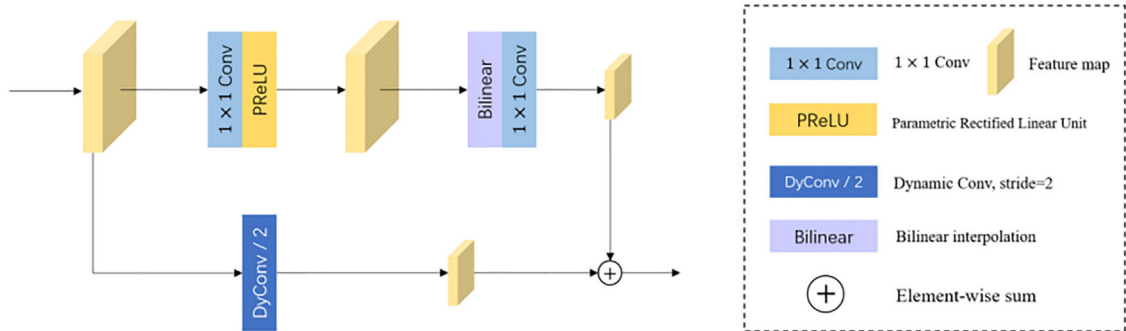


FIGURE 3 The structure of dynamic down-sampling module (DDM). One dynamic convolutional kernel used in DDM is synthesized with four traditional convolutional kernels and their corresponding weights are generated by sigmoid function

enriched features and ease information loss during the process of resolution reduction, dynamic convolution [68, 69], which is a lightweight plug-in with strong representation ability, is employed to build the dynamic down-sampling module (DDM). Meanwhile, hybrid attention module (HAUM) is also proposed for better interaction between cross-scale features. By considering both channel attention and spatial attention, HAUM can effectively absorb rich contextual information from low-resolution features and further guide accurate high-resolution reconstruction. Detailed descriptions about DDM and HAUM are given in the following paragraphs.

The operation of down-sampling is an effective way to enlarge the receptive field. Traditional methods using bilinear interpolation or pooling often cause serious distortion and fail to learn enriched features. We propose a lightweight down-

sampling subnetwork, namely dynamic down-sampling module (DDM). Its structure is illustrated in Figure 3. In each DDM, one branch containing a dynamic convolutional layer extracts features of interest and the other branch based on 1×1 convolution layers helps to enhance details of low-resolution features. Note that the dynamic convolution is an efficient plug-in and introduced into our module. Different from traditional convolution layers, its parameters change with the input so that it can significantly improve the representation ability. And the cost of additional runtime is very little. Two branches first work separately, and then the output features of each branch are aggregated for better feature extraction performance.

Considering that up-sampled features are inevitably inaccurate spatially, we propose a hybrid attention up-sampling module (HAUM), whose structure is shown in Figure 4. Motivated

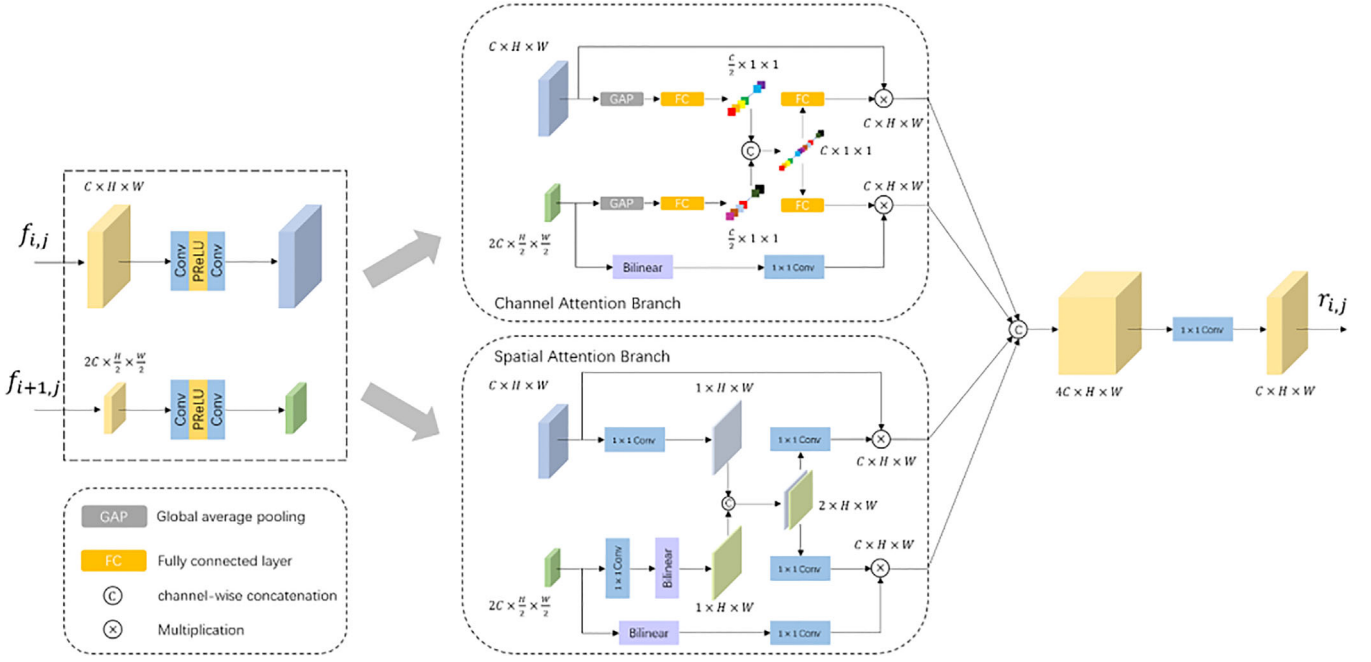


FIGURE 4 The structure of hybrid attention up-sampling module (HAUM). Based on dual attention [65], we further combine high-resolution features and low-resolution features. The multiplication operation in channel attention is channel-wise while it is element-wise in spatial attention branch

by the advances of dual attention, two branches are designed for channel attention and spatial attention, respectively.

Specifically, HAUM has two inputs (a high-resolution feature $f_{i,j}$ and a low-resolution feature $f_{i+1,j}$). After two convolutional layers with the kernel size of 3×3 , two-scale features are packed and fed into channel attention (CA) branch and spatial attention (SA) branch in parallel. The output of each branch will be concatenated. And the subsequent 1×1 convolution layer is used to reduce the number of channels and capture cross-channel information. Because HAUM generates a residual feature with the same size as high-resolution feature, we can further have:

$$f_{i,j+1} = f_{i,j} + r_{i,j} \quad (7)$$

$$r_{i,j} = \text{HAUM}(f_{i,j}, f_{i+1,j}) \quad (8)$$

where $\text{HAUM}(\cdot)$ denotes the operation of HAUM. $f_{i,j}$ is the j -th feature map in i -th branch in RHAB and $r_{i,j}$ is the output of HAUM.

CA branch and SA branch make the network pay more attention promoting informative information flow and suppressing relatively secondary features. They follow the similar design and introduce squeeze and excitation (SE) operations proposed in SENet into each branch. Take CA branch as an example, global average pooling (GAP) is applied to obtain channel-wise global information of both high-resolution and low-resolution features. They are respectively encoded into vectors with lower dimensions by fully connected (FC) layers to capture channel-wise dependencies. These two vectors are then concatenated and another two FC layers decode the concatenated result into

channel attention for low-resolution and high-resolution features. Finally, input features are multiplied by the generated channel attentions. In SA branch, spatial attention will be generated in a similar way, but we use 1×1 convolution to get spatial-wise global information instead of GAP.

Advantages of this design are obvious. First, it can focus on informative components with a comprehensive attention mechanism. So, the generated features are more task-sensitive and this can help to learn enriched representations. Besides, our HAUM takes both high-resolution and low-resolution features into consideration so that high-level contextual information can be integrated into high-resolution branch without sacrificing spatial accuracy. Therefore, “hybrid attention” has two following meanings:

- The hybrid of channel attention mechanism and spatial attention mechanism;
- The hybrid of features with high-resolution spatial accuracy and features with high-level contextual information.

4 | EXPERIMENTAL RESULTS

In this section, comprehensive experiments are implemented to verify the effectiveness of our HAN. First, we conduct experiments in three independent subtasks: low-light image enhancement, image super-resolution and image denoising. Quantitative and qualitative analysis are done on benchmark datasets. In order to make a fair comparison, we select some most widely used and representative benchmark datasets for each subtask. Second, based on LOL dataset [17], which is a low-light image

enhancement dataset containing 1000 synthetic image pairs, we further create a new dataset where images are corrupted with more complex degradation kernel. Our proposed HAN and some other existing state-of-the-art or classical networks are performed on this new dataset. Finally, we compare quantitative and visual results of our method and others to evaluate the performance in image quality enhancement.

4.1 | Datasets and metric

LOL dataset is used for low-light image enhancement task. The dataset is the first low-light real-world dataset with paired images. It contains 485 image pairs for training and 15 image pairs for testing.

SIDD [70], used in real-world image denoising, is a dataset containing image pairs captured by smart phones. In this dataset, the number of high-resolution training pairs is 320 while the number of cropped image pairs for testing is 1280.

In image super-resolution task, a famous dataset DIV2K consisting of 800 high-resolution images [71] is used for training in our experiments. Set5 [72], Set14 [73], B100 [74], Urban100 [75] and Manga109 [76] are used for testing. They are the most widely-used benchmark datasets for image super-resolution. In this task, only scale-factor 4 with bicubic degradation is taken into consideration. One reason of such choice is that the scale factor is often set to 2, 3, 4 or even higher in mainstream methods while many previous algorithms are only evaluated with the scale factor not greater than 4. Besides, the larger the scale factor, the more difficult it is to recover the lost details, which can better reflect the performance of networks. In addition, bicubic interpolation is the most widely-used degradation kernel in the field of image super-resolution. Therefore, it is convenient to make comparisons with other methods on this condition.

Image quality enhancement is a quite complex task, where low-quality images are corrupted by multiple degraded models. In order to obtain low-quality images and corresponding ground truths, we develop an Image Quality Enhancement Dataset (IQED) based on 1000 images from LOL dataset. These 1000 image pairs are initially used in low-light image enhancement. We first down-sample low-light images with bicubic kernel and then add Gaussian noise with noise level 15. In IQED, 100 images are randomly selected for testing while the left are used for training. The synthetic dataset IQED is published on <https://github.com/356056849/HAN>.

Metrics used for evaluation are peak signal to noise ratio (PSNR), structural similarity (SSIM), visual information fidelity (VIF) and multi-scale structural similarity (MS-SSIM). To be specific, PSNR measures the ration between the maximum power of the signal and the mean power of the noise. SSIM index is a kind of full reference metric that measures the structural similarity between two images. VIF is an index based on the mutual information between two images and MS-SSIM is an extension of SSIM which is closer to the human visual system. For each sub-task, in order to ensure the best performance of other methods, we directly collect the PSNR and SSIM results

from their corresponding papers. Due to the lack of other common metrics in these sub-tasks, we only compute the PSNR and SSIM indexes of our HAN. For image quality enhancement, all above metrics are used for evaluation because other methods and our HAN are trained from scratch on IQED dataset in this task.

4.2 | Implementation details

In our RHAB, there exist three branches where sizes of feature maps are different. In the first branch ($\times 1$ resolution), the number of filters is set to 64. While in the second branch ($\times \frac{1}{2}$ resolution) and the third branch ($\times \frac{1}{4}$ resolution), numbers of filters are set to 128 and 256, respectively. The number of RHABs b is 12 in low-light image enhancement and image denoising while it is 16 in image super-resolution. However, b is only 8 in the task of image quality enhancement.

In training stage, data augmentation, including random rotation and flip, is performed in our experiments. In all tasks, 144×144 patches are randomly extracted from training datasets. For image super-resolution, low-resolution images are generated with scale factor $\times 4$ while in the task of image quality enhancement, scale factors are $\times 2$, $\times 3$, and $\times 4$. Our network is trained with the optimizer Adam [77], and β_1 , β_2 and ϵ are set to 0.9, 0.999 and 10^{-8} , respectively. The learning rate is set to 2×10^{-4} initially and decreases to half every 1.5×10^5 iterations. Our experiments are conducted on a Geforce RTX 2080 Ti using Pytorch.

4.3 | Results and comparisons

In this section, we first compare our method with other existing algorithms on benchmark datasets. Then, both quantitative and visual comparisons are performed in image quality enhancement task. Note that results of all methods on IQED are implemented by us with the same training settings.

4.3.1 | Low-light image enhancement

Our network is trained and evaluated on LOL dataset. Best PSNR/SSIM values of the authors' HAN are reported for comparison with some other algorithms [2, 17, 18, 78, 79] quantitatively and results are recorded in Table 1. It can be observed that our HAN achieves very advanced performance.

In terms of PSNR, the evaluation result of our HAN is only 0.66 dB lower than that of MIRNet but outperforms all other methods. However, the SSIM of results generated by our HAN achieves the best score, which is 0.01 higher than state-of-the-art methods at least.

For visual comparisons, we present some results of our HAN and some other methods in Figure 5. Although the illumination of our output image may be less close to ground truth than that

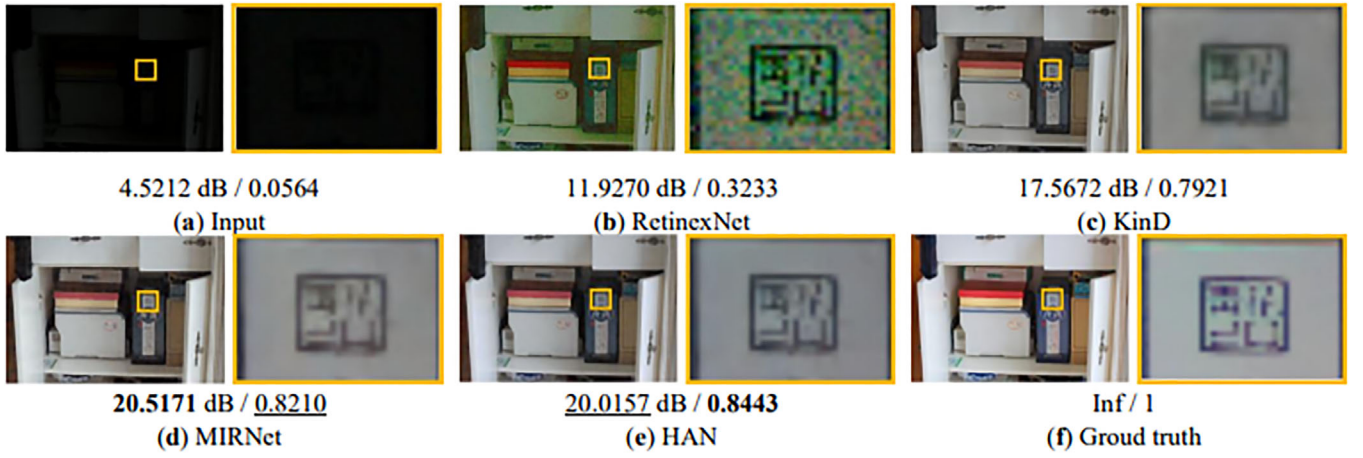


FIGURE 5 Visual comparisons on LOL datasets. (a) Low-light image; (b) RetinexNet [17]; (c) KinD [18]; (d) MIRNet [2]; (e) HAN; (f) Ground truth

TABLE 1 Quantitative comparisons of recent state-of-the-art methods for low-light image enhancement on LOL

Method	PSNR	SSIM
LIME [78]	16.76	0.56
RetinexNet [17]	16.77	0.56
GLAD [79]	19.72	0.70
KinD [18]	20.87	0.80
MIRNet [2]	24.14	<u>0.83</u>
HAN (Ours)	<u>23.48</u>	0.84

Best and second-best results are **highlighted** and underlined, respectively

generated by MIRNet, our algorithm can reconstruct images with higher clarity. And this is consistent with the trend of quantitative results.

4.3.2 | Image denoising

In this section, we will analyse experimental results for image denoising. Our network is trained and evaluated on SIDD datasets. Quantitative comparisons based on PSNR and SSIM are shown in Table 2. Our HAN achieves second-best performance among state-of-the-art methods. In terms of PSNR, our HAN is only 0.04 dB lower than the best one, while the SSIM gap is even smaller, which is 0.001.

Also, visual results are shown in Figure 6, together with state-of-the-art methods. It can be observed that our HAN can effectively remove real noise and produce very competing results compared to recent best denoiser.

4.3.3 | Image super-resolution

We compare our proposed network with 7 other image super-resolution algorithms: Bicubic, SRCNN [39], VDSR [29], LapSRN [66], SRDenseNet [28], D-DBPN [83] and RDN [3]

TABLE 2 Quantitative comparisons of recent state-of-the-art methods for image denoising on SIDD

Method	PSNR	SSIM
DnCNN [23]	23.66	0.583
BM3D [80]	25.65	0.685
CBDNet [22]	30.78	0.754
RIDNet [21]	38.71	0.914
VDN [50]	39.28	0.909
CycleISP [6]	39.52	0.957
DRDN [81]	39.60	0.940
MIRNet [2]	39.72	0.959
HAN (ours)	<u>39.68</u>	<u>0.958</u>

Best and second-best results are **highlighted** and underlined, respectively

on Set5, Set14, B100, Urban100 and Manga109. Quantitative results are collected in Table 3. Our HAN achieves the second-best quantitative performance on most items and results are very close to the best. The average PSNR/SSIM values of our results on Set5 are only 0.01 dB and 0.0001 lower than that generated by RDN. However, there is still a significant gap between our HAN and RDN on the Manga109 dataset.

Visual comparisons in Figure 7 shows that our HAN can reconstruct images with rich details and sharp edges. It can be observed that most of methods cause noticeable and unpleasant artifacts. Although the performance of our method is not as good as RDN in terms of quantitative results, it still achieves very good performance visually. For human eyes, the image generated by our HAN is only slightly different from that generated by RDN.

4.3.4 | Image quality enhancement

Image quality enhancement aims to restore low-light, low-resolution and noisy images. All experiments in this section



FIGURE 6 Visual comparisons on SIDD datasets. (a) Noisy image; (b) CBDNet [22]; (c) RIDNet [21]; (d) VDN [50]; (e) CycleISP [6]; (f) MIRNet [2]; (g) HAN; (h) Ground truth

TABLE 3 Quantitative comparisons of recent state-of-the-art methods for image super-resolution (scale factor $\times 4$) on Set5, Set14, B100, Urban100 and Manga109

Method	Set5 PSNR/SSIM	Set14 PSNR/SSIM	B100 PSNR/SSIM	Urban100 PSNR/SSIM	Manga109 PSNR/SSIM
Bicubic [82]	24.82/0.8104	26.00/0.7027	25.96/0.6675	23.14/0.6577	24.89/0.7866
SRCNN [39]	30.48/0.8628	27.50/0.7513	26.90/0.7101	24.52/0.7221	27.58/0.8555
VDSR [29]	31.35/0.8830	28.02/0.7680	27.29/0.0726	25.18/0.7540	28.83/0.8870
LapSRN [66]	31.54/0.8850	28.19/0.7720	27.32/0.7270	25.21/0.7560	29.09/0.8900
SRDenseNet [28]	32.02/0.8930	28.50/0.7780	27.53/0.7337	26.05/0.7819	N/A/N/A
D-DBPN [83]	32.47/0.8980	<u>28.82/0.7860</u>	<u>27.72/0.7400</u>	26.38/0.7946	30.91/0.9137
RDN [3]	32.61/0.8999	28.93/0.7894	27.80/0.7436	26.85/0.8089	31.45/0.9187
HAN (ours)	<u>32.60/0.8997</u>	28.78/ <u>0.7869</u>	<u>27.71/0.7411</u>	<u>26.67/0.8025</u>	<u>31.15/0.9168</u>

Best and second-best results are **highlighted** and underlined, respectively

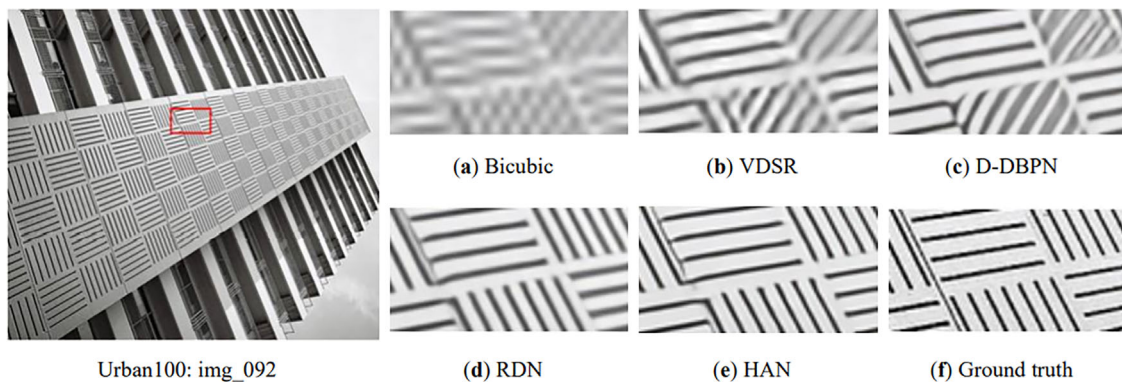


FIGURE 7 Visual comparisons on img_092 in Urban100 dataset. (a) Bicubic; (b) VDSR [29]; (c) D-DBPN [83]; (d) RDN [3]; (e) HAN; (f) Ground truth

TABLE 4 Quantitative comparisons of recent state-of-the-art methods for image quality enhancement on IQED

Method	Scale	PSNR (dB)	SSIM	VIF	MS-SSIM
BM3D [80]+Bicubic [82]+LIME [78]	×2	15.39	0.4856	0.2350	0.6720
RetinexNet [17]	×2	15.34	0.2944	0.1489	0.5495
U-net [31]	×2	18.00	0.5757	0.2412	0.7499
DnCNN [23]	×2	18.40	0.5278	0.2415	0.7403
ESPCN [31]	×2	18.44	0.5628	0.2410	0.7338
RDN [3]	×2	19.59	0.6605	<u>0.3299</u>	0.8216
MIRNet [2]	×2	<u>20.11</u>	0.6462	0.3197	<u>0.8280</u>
HAN (ours)	×2	20.51	<u>0.6592</u>	0.3364	0.8370
BM3D [80]+Bicubic [82]+LIME [78]	×3	15.21	0.4856	0.2064	0.6204
RetinexNet [17]	×3	14.48	0.2771	0.1251	0.5138
U-net [31]	×3	17.49	0.5271	0.2056	0.7019
DnCNN [23]	×3	17.95	0.5214	0.2243	0.7242
ESPCN [31]	×3	18.06	0.5369	0.2275	0.7093
RDN [3]	×3	18.53	0.6163	<u>0.3085</u>	0.7897
MIRNet [2]	×3	<u>19.86</u>	<u>0.6209</u>	0.3082	<u>0.8035</u>
HAN (ours)	×3	19.88	0.6316	0.3214	0.8141
BM3D [80]+Bicubic [82]+LIME [78]	×4	14.97	0.4444	0.1874	0.5736
RetinexNet [17]	×4	14.38	0.2572	0.1182	0.4920
U-net [31]	×4	17.29	0.5107	0.1882	0.6871
DnCNN [23]	×4	17.53	0.5050	0.2001	0.6944
ESPCN [31]	×4	17.51	0.5181	0.2144	0.6814
RDN [3]	×4	18.18	0.5827	0.2830	0.7574
MIRNet [2]	×4	<u>19.18</u>	<u>0.5920</u>	0.3024	0.7775
HAN (ours)	×4	19.42	0.5933	<u>0.2831</u>	<u>0.7727</u>

Best and second-best results are **highlighted** and underlined, respectively

are performed in our IQED dataset. For a fair comparison, we implement other recent state-of-the-art algorithms [2, 3, 31] and pioneer methods [17, 23, 78, 80, 82]. They are trained and evaluated with the same settings as ours. Note that for models that do not contain an upscale module, the input image is first up-sampled to target size with bilinear interpolation. Table 4 shows quantitative comparisons evaluated for scale factors ×2, ×3 and ×4. In the task of image quality enhancement, our HAN outperforms most of other previous methods quantitatively. To be specific, for scale factor ×2, ×3 and ×4, our algorithm achieves the best performance on 3, 4 and 2 metrics, respectively. The evaluation result on PSNR of our method is higher than all those of other methods. In terms of other metrics, our method achieves better performance in most cases. Quantitatively, it can be regarded that our HAN performs the best on IQED among other state-of-the-art methods.

In Figure 8, we show visual comparisons on IQED with the scale factor ×2. It can be observed that U-net produces an unpleasant bright spot at the center of the enhanced image and it fails to remove the noise. By comparing results generated by different methods, it can be found that the image enhanced by our

method is cleaner and its colour is closer to the ground truth. Another visual comparison is given in Figure 9. Less artifacts are brought about by our algorithm compared to other state-of-the-art methods, despite the problems of colour deviation, noticeable distortion, and unnaturalness due to extreme degradations.

In order to compare the complexity of our proposed method with others, we calculate the number of parameters and average runtime under different scale factors. For a fair comparison, only CNN-based algorithms can be performed on GPUs with the same setting. Results are shown in Table 5. It can be found that our network has the largest number of parameters. Even so, it is about 10 ms faster than MIRNet, which has the second largest model size. It indicates that our model is more efficient than MIRNet. The reasons why MIRNet and our HAN are slower than others is that these two methods maintain feature maps of different resolution for the sake of spatial accuracy and they have much more parameters. Hence, this compromise seems to be inevitable. Since the focus of our method is to improve the ability of the network to deal with complex degradation kernels, it can be considered that such compromise in efficiency is acceptable.

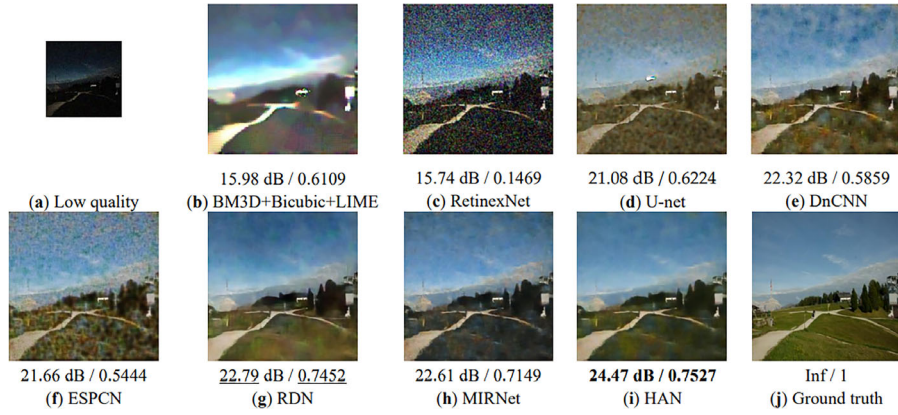


FIGURE 8 Visual comparisons of r1cc0e4e1t on IQED dataset with the scale factor $\times 2$. (a) Low quality; (b) BM3D [80]+Bicubic[82]+LIME [78]; (c) RetinexNet [17]; (d) U-net [31]; (e) DnCNN [23]; (f) ESPCN [31]; (g) RDN [3]; (h) MIRNet [2]; (i) HAN; (j) Ground truth

FIGURE 9 Visual comparisons of r089cae91t on IQED dataset with the scale factor $\times 4$. (a) Low quality; (b) BM3D [80]+Bicubic[82]+LIME [78]; (c) RetinexNet [17]; (d) U-net [31]; (e) DnCNN [23]; (f) ESPCN [31]; (g) RDN [3]; (h) MIRNet [2]; (i) HAN; (j) Ground truth

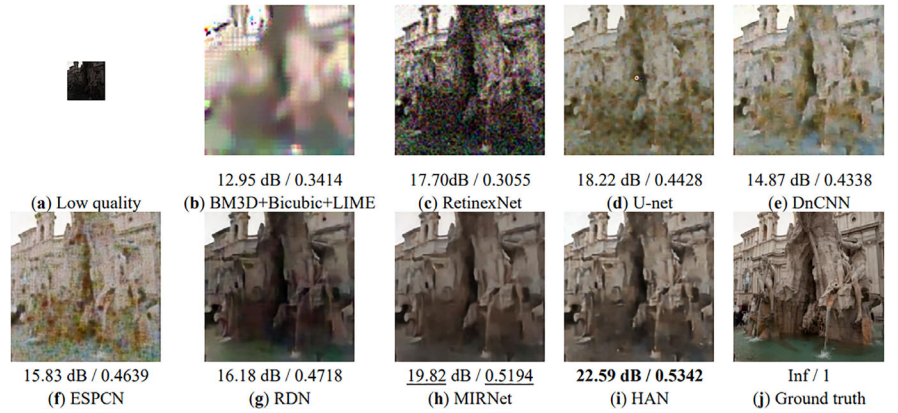


TABLE 5 The number of parameters and runtime speed of different methods

	parameters (M)	$\times 2$ (ms)	$\times 3$ (ms)	$\times 4$ (ms)
RetinexNet	0.44	3.11	3.11	3.10
DnCNN	0.56	3.93	3.89	3.91
ESPCN	0.03	0.45	0.43	0.43
U-net	31.04	5.37	5.38	5.30
RDN	22.27	23.96	16.35	16.54
MIRNet	<u>31.78</u>	67.68	67.72	65.44
HAN (ours)	35.99	<u>57.45</u>	<u>57.25</u>	<u>56.95</u>

The runtime means the average inference time of 100 image patches with the size of 144×144 . Biggest/slowest results are **highlighted** and second-biggest/slowest results are underlined

5 | ABLATION STUDY

In order to study the effectiveness of each module proposed in our HAN, we perform the ablation experiments for image quality enhancement with scale factor $\times 4$. We replace our DDMs with a convolutional layer with stride 2 to validate the impacts of this module. Similarly, 1×1 convolution and the bilinear inter-

TABLE 6 Results of the ablation study for image quality enhancement ($\times 4$)

DDM	HAUM	PSNR (dB)/SSIM
X	X	18.22/0.5804
✓	X	19.36/0.5858
X	✓	19.41/0.5911
✓	✓	19.42/0.5933

“X” and “✓” indicate whether the module is used

polation are used to replace HAUMs. Results of the ablation study are shown in Table 6. It can be observed that both DDM and HAUM can significantly improve the performance of our HAN. When they work together, our HAN achieves the best quantitative results.

6 | DISCUSSION

The novel architecture design of our HAN leads to a good trade-off between high-resolution spatial accuracy and rich contextual information. In order to better illustrate the innovations

of our network, we have made some simple comparisons with other methods:

- Comparison to U-net++ [67]. We draw lessons from nested pathways designed in U-net++ and introduce this structure into our RHAB. However, in U-net++, the way of feature fusion is concatenation while it is sum in our RHAB.

In addition, in RHAB, down-sampling and up-sampling are based on DDM and HAUM, respectively. In U-net++, they are based on bilinear interpolation, convolution, or transposed convolution, though.

- Comparison to dual attention unit (DAU). Like dual attention, our HAUM also has two branches, respectively for spatial attention and channel attention. In both two methods, the output feature maps of two branches are concatenated together. The difference between DAU and our HAUM is that the input of the former is one feature tensor, while two feature tensors with different sizes are fed into HAUM instead. DAU tries to learn spatial-aware and channel-aware representations, but HAUM further combines multi-scale features to balance spatial accuracy and contextual information.

7 | CONCLUSIONS

Existing methods in image quality enhancement mainly operate on single resolution or are based on encoder-decoder structure. Besides, in different subtasks, the design of networks follows different pipelines. Thus, we propose a novel architecture which is a unified framework in image quality enhancement, to solve above problems. The key component of it consists of multiple branches, where multi-scale features fuse and separate repeatedly with the help of proposed dynamic down-sampling modules and hybrid attention up-sampling modules, to better promote the flow of global information and local information. Further, experimental results demonstrate that our method can achieve comparable performance against state-of-the-art algorithms in independent subtask and it outperforms other reimplemented state-of-the-art methods in the task of image quality enhancement. Therefore, it can be concluded that there are at least two advantages of our method. One advantage is that for different tasks, only the training data need to be changed to achieve good performance. Another advantage is that our method can better adapt to the real world because the low-quality images in the real scene are often corrupted by multiple degradation kernels. It is believed that proposed algorithm can be used as a pre-processing method of downstream high-level vision tasks, help professionals of photography and film to improve the quality of captured photos or videos, and reduce the difficulty of recognizing contents from monitoring cameras.

However, there still exist some problems. First, our data is artificially synthesized while there is still a gap between the synthesized data and the real scene. Directly collecting a large num-

ber of paired training data in the real world is a solution but it is quite time-costing. In addition, the generated images still lack sufficient visual quality. In future work, to tackle the first problem, we would like to collect some unpaired training data and explore to train the model in an unsupervised or semi-supervised way. To solve the second problem, introducing other loss functions, such as perceptual loss and discriminative loss, into our method to generate more photorealistic images could be another future direction.

ACKNOWLEDGEMENTS

This work is supported by the National Key Research and Development Program of China (No. 2020YFB1406800).

AUTHOR CONTRIBUTIONS

Jiachen Wang: formal analysis; investigation; methodology; software; validation; writing – original draft. **Yingyun Yang:** conceptualization; funding acquisition; project administration; resources; supervision; writing-review and editing. **Yan Hua:** writing – review and editing.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Jiachen Wang  <https://orcid.org/0000-0002-3252-375X>

REFERENCES

1. Zamir, S.W., Arora, A., Khan, S., et al.: Multi-stage progressive image restoration. arXiv:2102.02808 (2021)
2. Zamir, S.W., Arora, A., Khan, S., et al.: Learning enriched features for real image restoration and enhancement. arXiv:2003.06792 (2020)
3. Zhang, Y., Tian, Y., Kong, Y., et al.: Residual dense network for image restoration. *IEEE Trans. Pattern Anal. Mach. Intell.* 43(7), 2480–2495 (2020)
4. Wan, Z., Zhang, B. & Chen, D. et al.: Bringing old photos back to life. In: *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2747–2757. Seattle, WA (2020)
5. Abu-Hussein, S., Tirer, T., Chun, S.Y., et al.: Image restoration by deep projected GSURE. arXiv:2102.02485 (2021)
6. Zamir, S.W., Arora, A. & Khan, S. et al.: CycleISP: Real image restoration via improved data synthesis. In: *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2696–2705. Seattle, WA (2020)
7. Pan, X., Zhan, X. & Dai, B. et al.: Exploiting deep generative prior for versatile image restoration and manipulation. In: *Proc. European Conference on Computer Vision*, pp. 262–277. Glasgow, UK (2020)
8. Zhang, H., Sun, L., Wu, L., et al.: DuGAN: An effective framework for underwater image enhancement. *IET Image Process.* 15(9), 2010–2019 (2021)
9. Li, F., Zheng, J., Zhang, Y.: Generative adversarial network for low-light image enhancement. *IET Image Process.* 15(7), 1542–1552 (2021)
10. Li, S., Qin, B., Xiao, J., et al.: Multi-channel and multi-model-based autoencoding prior for grayscale image restoration. *IEEE Trans. Image Process.* 29, 142–156 (2020)
11. Jin, Z., Iqbal, M.Z., Bobkov, D., et al.: A flexible deep CNN framework for image restoration. *IEEE Trans. Multimed.* 22(4), 1055–1068 (2020)

12. Yu, Y., Liu, M., Feng, H., et al.: Split-attention multiframe alignment network for image restoration. *IEEE Access* 8, 39254–39272 (2020)
13. Huang, H., Schiopu, I., Munteanu, A.: Macro-pixel-wise CNN-based filtering for quality enhancement of light field images. *Electron. Lett.* 56, 1413–1416 (2020)
14. Singh, N., Bhandari, A.K.: Image contrast enhancement with brightness preservation using an optimal gamma and logarithmic approach. *IET Image Process.* 14, 794–805 (2019)
15. Land, E., McCann, J.: Lightness and retinex theory. *J. Opt. Soc. Am.* 61(1), 1–11 (1971)
16. Shen, L., Yue, Z., Feng, F., et al.: MSR-net: Low-light image enhancement using deep convolutional network. *arXiv:1711.02488* (2017)
17. Wei, C., Wang, W. & Yang, W. et al.: Deep retinex decomposition for low-light enhancement. In: *Proc. British Machine Vision Conference*, Newcastle, UK (2018)
18. Zhang, Y., Zhang, J. & Guo, X.: Kindling the darkness: A practical low-light image enhancer. In: *Proc. 27th ACM International Conference on Multimedia*, pp. 1632–1640. Nice, France (2019)
19. Wang, R., Zhang, Q. & Fu, C.W. et al.: Underexposed photo enhancement using deep illumination estimation. In: *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6849–6857. Long Beach, CA (2019)
20. Jiang, Y., Gong, X., Liu, D., et al.: EnlightenGAN: Deep light enhancement without paired supervision. *IEEE Trans. Image Process.* 30, 2340–2349 (2021)
21. Anwar, S. & Barnes, N.: Real image denoising with feature attention. In: *Proc. IEEE/CVF International Conference on Computer Vision*, pp. 3155–3164. Seoul, South Korea (2019)
22. Shi, G., Yan, Z. & Kai, Z. et al.: Toward convolutional blind denoising of real photographs. In: *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1712–1722. Long Beach, CA (2019)
23. Kai, Z., Zuo, W., Chen, Y., et al.: Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising. *IEEE Trans. Image Process.* 26(7), 3142–3155 (2016)
24. Zhang, K., Zuo, W., Zhang, L.: FFDNet: Toward a fast and flexible solution for CNN-based image denoising. *IEEE Trans. Image Process.* 27, 4608–4622 (2018)
25. Li, J., Fang, F. & Mei, K. et al.: Multi-scale residual network for image super-resolution. In: *Proc. European Conference on Computer Vision*, pp. 517–532. Munich, Germany (2018)
26. Lim, B., Son, S. & Kim, H. et al.: Enhanced deep residual networks for single image super-resolution. In: *Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 136–144. Honolulu, HI (2017)
27. Zhang, Y., Tian, Y. & Kong, Y. et al.: Residual dense network for image super-resolution. In: *Proc. Conference on Computer Vision and Pattern Recognition*, pp. 2472–2481. Salt Lake City, UT (2018)
28. Tong, T., Li, G. & Liu, X. et al.: Image super-resolution using dense skip connections. In: *Proc. IEEE International Conference on Computer Vision*, pp. 4799–4807. Venice, Italy (2017)
29. Zhang, Y., Li, K. & Li, K. et al.: Image super-resolution using very deep residual channel attention networks. In: *Proc. European Conference on Computer Vision*, pp. 286–301. Munich, Germany (2018)
30. Shi, W., Caballero, J. & Huszar, F. et al.: Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1874–1883. Las Vegas, NV (2016)
31. Ronneberger, O., Fischer, P. & Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: *Proc. International Conference on Medical Image Computing and Computer-assisted Intervention*, pp. 234–241. Munich, Germany (2015)
32. Long, J., Shelhamer, E. & Darrell, T.: Fully convolutional networks for semantic segmentation. In: *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3431–3440. Boston, MA (2015)
33. Lore, K.G., Akinlayo, A., Sarkar, S.: LLNet: A deep autoencoder approach to natural low-light image enhancement. *Pattern Recognit.* 61, 650–662 (2017)
34. Chen, C., Chen, Q. & Xu, J. et al.: Learning to see in the dark. In: *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3291–3300. Salt Lake City, UT (2018)
35. Mao, X.J., Shen, C., Yang, Y.B.: Image restoration using convolutional auto-encoders with symmetric skip connections. *arXiv:1606.08921* (2016)
36. Abbasi, A., Monadjem, A., Fang, L., et al.: Three-dimensional optical coherence tomography image denoising through multi-input fully-convolutional networks. *Comput. Biol. Med.* 108, 1–8 (2019)
37. Couturier, R., Perrot, G. & Salomon, M.: Image denoising using a deep encoder-decoder network with skip connections. In: *Proc. International Conference on Neural Information Processing*, pp. 554–565. Montréal, Canada (2018)
38. Tao, L., Zhu, C. & Xiang, G. et al.: LLCNN: A convolutional neural network for low-light image enhancement. In: *Proc. IEEE Visual Communications and Image Processing*, pp. 1–4. St. Petersburg, FL (2017)
39. Dong, C., Loy, C.C., He, K.: Image super-resolution using deep convolutional networks. *IEEE Trans. Pattern Anal. Mach. Intell.* 38(2), 295–307 (2015)
40. Ignatov, A., Kobyshev, N. & Timofte, R. et al.: DSLR-Quality photos on mobile devices with deep convolutional networks. In: *Proc. IEEE International Conference on Computer Vision*, pp. 3277–3285. Venice, Italy (2017)
41. He, K., Zhang, X. & Ren, S. et al.: Deep residual learning for image recognition. In: *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778. Las Vegas, NV (2016)
42. Szegedy, C., Liu, W. & Jia, Y. et al.: Going deeper with convolutions. In: *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–9. Boston, MA (2015)
43. Jobson, D.J., Rahman, Z., Woodell, G.A.: A multiscale retinex for bridging the gap between color images and the human observation of scenes. *IEEE Trans. Image Process.* 6(7), 965–976 (1997)
44. Li, M., Zhou, D., Nie, R., et al.: AMBCR: Low-light image enhancement via attention guided multi-branch construction and Retinex theory. *IET Image Process.* 15(9), 2020–2038 (2021)
45. Yang, J., Xu, Y., Yue, H., et al.: Low-light image enhancement based on Retinex decomposition and adaptive gamma correction. *IET Image Process.* 15(5), 1189–1202 (2020)
46. Burger, H.C., Schuler, C.J. & Harmeling, S.: Image denoising: Can plain neural networks compete with BM3D? In: *Proc. IEEE Conference Computer Vision and Pattern Recognition*, pp. 2392–2399. Providence, RI (2012)
47. Xie, J., Xu, L. & Chen, E.: Image denoising and inpainting with deep neural networks. In: *Proc. Advances in Neural Information Processing Systems*, pp. 341–349. Lake Tahoe, NV (2012)
48. Ioffe, S. & Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: *Proc. International Conference on Machine Learning*, pp. 448–456. Lille, France (2015)
49. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. *arXiv:1511.07122* (2015)
50. Yue, Z., Yong, H., Zhao, Q., et al.: Variational denoising network: Toward blind noise modeling and removal. *arXiv:1908.11314* (2019)
51. Dong, C., Loy, C.C. & He, K.: Learning a deep convolutional network for image super-resolution. In: *Proc. European Conference on Computer Vision*, pp. 184–199. Zurich, Switzerland (2014)
52. Johnson, J., Alahi, A. & Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: *Proc. European Conference on Computer Vision*, pp. 694–711. Amsterdam, The Netherlands (2016)
53. Kim, J., Lee, J.K. & Lee, K.M.: Accurate image super-resolution using very deep convolutional networks. In: *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1646–1654. Las Vegas, NV (2016)
54. Huang, G., Liu, Z. & Van Der Maaten, L. et al.: Densely connected convolutional networks. In: *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4700–4708. Honolulu, HI (2017)
55. Fang, F., Li, J., Zeng, T.: 'Soft-edge assisted network for single image super-resolution. *IEEE Trans. Image Process.* 29, 4656–4668 (2020)
56. Lan, R., Sun, L., Liu, Z., et al.: Cascading and enhanced residual networks for accurate single-image super-resolution. *IEEE Trans. Cybern.* 51(1), 115–125 (2021)

57. Sun, K., Xiao, B. & Liu, D. et al.: Deep high-resolution representation learning for human pose estimation. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5693–5703. Long Beach, CA (2019)
58. Huo, Z., Jin, H., Qiao, Y., et al.: Deep high-resolution network with double attention residual blocks for human pose estimation. *IEEE Access* 8, 224947–224957 (2020)
59. Wang, F., Piao, S., Xie, J.: CSE-HRNet: A context and semantic enhanced high-resolution network for semantic segmentation of aerial imagery. *IEEE Access* 8, 182475–182489 (2020)
60. Zhang, R.: Making convolutional networks shift-invariant again. In: Proc. International Conference on Machine Learning, pp. 7324–7334. Long Beach, California (2019)
61. Qin, J., Sun, X., Yan, Y., et al.: Multi-resolution space-attended residual dense network for single image super-resolution. *IEEE Access* 8, 40499–40511 (2020).
62. Hu, J., Shen, L. & Sun, G.: Squeeze-and-excitation networks. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141. Salt Lake City, UT (2018)
63. Woo, S., Park, J. & Lee, J.Y. et al.: Cbam: Convolutional block attention module. In: Proc. European Conference on Computer Vision, Munich, Germany (2018) pp. 3–19
64. Tian, Y., Wang, Y., Yang, L., et al.: CANet: Concatenated attention neural network for image restoration. *IEEE Signal Process. Lett.* 27, 1615–1619 (2020)
65. Fu, J., Liu, J. & Tian, H. et al.: Dual attention network for scene segmentation. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3146–3154. Long Beach, CA (2019)
66. Lai, W.S., Huang, J.B. & Ahuja, N. et al.: Deep laplacian pyramid networks for fast and accurate super-resolution. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 624–632. Honolulu, HI (2017)
67. Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., et al.: Unet++: A nested U-net architecture for medical image segmentation. In: Proc. Deep learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, pp. 3–11. Granada, Spain (2018)
68. Chen, Y., Dai, X. & Liu, M. et al.: Dynamic convolution: Attention over convolution kernels. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11030–11039. Seattle, WA (2020)
69. Zhang, Y., Zhang, J., Wang, Q., et al.: Dynet: Dynamic convolution for accelerating convolutional neural networks. *arXiv:2004.10694* (2020)
70. Abdelhamed, A., Lin, S. & Brown, M.S. et al.: A high-quality denoising dataset for smartphone cameras. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 1692–1700. Salt Lake City, UT (2018)
71. Timofte, R., Agustsson, E. & Van Gool, L. et al.: Ntire 2017 challenge on single image super-resolution: Methods and results. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 114–125. Honolulu, HI (2017)
72. Bevilacqua, M., Roumy, A. & Guillemot, C. et al.: Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In: Proc. British Machine Vision Conference, Surrey, UK (2012)
73. Zeyde, R., Elad, M. & Protter, M.: On single image scale-up using sparse-representations. In: Proc. International Conference on Curves and Surfaces, pp. 711–730. Avignon, France (2010)
74. Martin, D., Fowlkes, C. & Tai, D. et al.: A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: Proc. IEEE International Conference on Computer Vision, pp. 416–423. Vancouver, Canada (2001)
75. Huang, J.B., Singh, A. & Ahuja, N.: Single image super-resolution from transformed self-exemplars. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 5197–5206. Boston, MA (2015)
76. Matsui, Y., Ito, K., Aramaki, Y., et al.: Sketch-based manga retrieval using manga109 dataset. *Multimed. Tools Appl.* 76(20), 21811–21838 (2017)
77. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. *arXiv:1412.6980* (2014)
78. Guo, X., Li, Y., LING, H.: LIME: Low-light image enhancement via illumination map estimation. *IEEE Trans. Image Process.* 26(2), 982–993 (2016)
79. Wang, W., Wei, C. & Yang, W. et al.: GLADNet: Low-light enhancement network with global awareness. In: Proc. IEEE International Conference on Automatic Face & Gesture Recognition, pp. 751–755. Xi'an, China (2018)
80. Dabov, K., Foi, A., Katkovnik, V., et al.: Image denoising by sparse 3-D transform-domain collaborative filtering. *IEEE Trans. Image Process.* 16(8), 2080–2095 (2007)
81. Song, Y., Zhu, Y., Du, X.: Dynamic residual dense network for image denoising *Sensors* 19(17), 3809 (2019).
82. Keys, R.: Cubic convolution interpolation for digital image processing. *IEEE Trans. Acoust. Speech Signal Process.* 29(6), 1153–1160 (1981)
83. Haris, M., Shakhnarovich, G. & Ukita, N.: Deep back-projection networks for super-resolution. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 1664–1673. Salt Lake City, UT (2018)

How to cite this article: Wang, J., Yang, Y., Hua, Y.: Image quality enhancement using hybrid attention networks. *IET Image Process.* 16, 521–534 (2022).
<https://doi.org/10.1049/ipr2.12368>