doi:10.3772/j. issn. 1006-6748. 2022. 02. 004

Non-identical residual learning for image enhancement via dynamic multi-level perceptual loss[①]

HU Ruiguang(胡瑞光), HUANG Li[②]
(Beijing Aerospace Automatic Control Institute, Beijing 100854, P. R. China)

Abstract

Residual learning based deep generative networks have achieved promising performance in image enhancement. However, due to the large color gap between a low-quality image and its high-quality version, the identical mapping in conventional residual learning cannot explore the elaborate detail differences, resulting in color deviations and texture losses in enhanced images. To address this issue, an innovative non-identical residual learning architecture is proposed, which views image enhancement as two complementary branches, namely a holistic color adjustment branch and a fine-grained residual generation branch. In the holistic color adjustment, an adjusting map is calculated for each input low-quality image, in order to regulate the low-quality image to the high-quality representation in an overall way. In the fine-grained residual generation branch, a novel attention-aware recursive network is designed to generate residual images. This design can alleviate the overfitting problem by reusing parameters and promoting the network's adaptability for different input conditions. In addition, a novel dynamic multi-level perceptual loss based on the error feedback ideology is proposed. Consequently, the proposed network can be dynamically optimized by the hybrid perceptual loss provided by a well-trained VGG, so as to improve the perceptual quality of enhanced images in a guided way. Extensive experiments conducted on publicly available datasets demonstrate the state-of-the-art performance of the proposed method.

Key words: image enhancement, deep residual network, adversarial learning

0 Introduction

Image enhancement, as a classical computer vision task, aims at recovering high-quality image from its low-quality version. High-quality images should have abundant color, clear texture, and satisfactory perception, etc.. It is an important task that can facilitate various industrial communities, e. g. satellite[1], medical, and 4K television[2]. Many traditional enhancement methods including Gaussian smoothing, and bilateral filtering have been proposed without supervised information. With the flourish of deep neural networks, convolutional neural networks (CNNs) have shown the powerful capability in image enhancement by learning pairwise training patches. Some existing methods mainly focus on solving image enhancement problem from specific aspects, such as enhancing illumination, adjusting contrast, and denoising.

It can be noticed that low-quality images and their high-quality targets have great similarity in contents, thus their detail differences, i. e. texture, edge and color recovery are important for image enhancement. Consequently, residual learning has become a successful method to excavate those details by building an identical mapping from low-quality to high-quality images. Later, generative adversarial networks (GANs) based image enhancement frameworks are proposed. They adopt deep residual network as generative model for enhancing low-quality images, and take multiple loss function, e. g. a perceptual loss and an adversarial loss, to optimize network for promoting visual quality. However, those methods still remains three deficiencies. (1) A low-quality image and its high-quality version exist large gaps in holistic color. The identical mapping in the residual learning cannot force generative models to accurately capture the detailed information. (2) Generative models usually have large number of parameters, causing great storage cost and rising the risk of overfitting. (3) Although one or multi-level perceptual losses are widely applied for network optimization, the loss weight allocated to each level are

[①] Supported by the National Natural Science Foundation of China (No. 62172035).

[②] To whom correspondence should be addressed. E-mail: 13611007479@163.com.

Received on June 12, 2021

fixed, resulting in unpleasant artifacts or unfavorable color representations in enhanced images.

To address the above-mentioned issues, non-identical residual learning is first considered to adjust low-quality images to high-quality style. Hence, a novel image enhancement framework is proposed, which consists of two complementary branches: holistic color adjustment and fine-grained residual generation. In the fine-grained residual generation branch, recursive structures are employed to construct the proposed network with less parameters meanwhile alleviate overfitting. However, the feature representations are still limited due to model capacity, and it lacks flexibility to adapt to different image scenes. Consequently, a lightweight attention-aware recursive network is proposed. It is composed of fully multi-scale feature extraction to extract more representative primary features, and a recursive convolutional function, which collocates multi-level channel-wise attention to promote the flexibility of the network by dynamically excavating color information. The holistic color adjustment can adjust global information and facilitate the generative network to learn local details. It is tried to compute the overall residuals between low-quality images and high-quality images. Then, an adjusting map is estimated for input low-quality images adaptively. Accordingly, low-level feature maps extracted from a well-trained network have abundant color information, while the extracted high-level feature maps contain more spatial and texture information. Optimizing single one-level perceptual loss cannot comprehensively promote enhanced quality. Therefore, a multi-level perceptual loss is considered to comprehensively optimize the proposed network. However, the loss weight of each level cannot be easily determined, and it lacks of flexibility during the training process. Consequently, a dynamic multi-level perceptual loss is introduced for optimization based on the error feedback. Detailedly, feature contents of high-quality and enhanced images are extracted from max-pooling layers of VGG16, and content errors between high-quality and enhanced features are computed. According to the value of the errors, a weight is decided for perceptual loss of each level. Thus, enhanced images will have rational color representations and textures.

In summary, the main contributions of this paper are as follows.

(1) A novel non-identical residual learning frame-work is tailored for image enhancement, in which an adjusting map is carefully computed to adjust global color to high-quality target.

(2) A novel attention-aware recursive network is proposed to adaptively enhance residual details according to input low-quality images.

(3) An innovative dynamic multi-level perceptual loss (DPL) is presented to approximate color representation of high-quality images, hence promoting perceptual effect in a more comprehensive way.

(4) Extensive experiments on publicly available dataset show the state-of-the-art performance of the proposed method, both quantitatively and qualitatively.

The rest of paper is organized as follows. Section 1 overviews related work. Section 2 describes the enhancement architecture. Experimental results and their analysis are presented in Section 3. Section 4 concludes this paper.

1 Related work

1.1 Image enhancement

The pioneer image enhancement work often concentrate on improving image contrast, such as histogram equalization (HE) and its variants bi-HE. Ref. [3] proposed a low-light image enhancement method by estimating illumination maps. However, those methods do not use external information and the performance of them is usually inadequate and limited. An external example-based approach was proposed for low-light image enhancement in Ref. [4], which adopts an auto-encoder to learn a mapping function. Ref. [5] proposed a unified image enhancement framework, which combines learning based methods with reconstruction based methods. Some work enhance images in specific conditions, e. g. hyper-spectral image and underwater image.

Recent years, CNNs show promising performance in many image enhancement sub-tasks, e. g. image super-resolution, image denosing[6] and image colorization. In Ref. [7], a reconstruction-based pairwise depth dataset for depth image enhancement was proposed. CNNs for weakly illuminated image enhancement was proposed in Ref. [3]. Deep residual learning was proposed in Ref. [8], and it showed effectiveness for deep network construction. However, those deep networks significantly increase the number of parameters and the overfitting problem is highly likely. Recursive structures have become an effective way to relieve overfitting for the less parameters. Ref. [9] proposed DRRN that combines residual learning for easy training by a 52-layers network, showing the promising performance in image super-resolution. Employing the recursive structure is tried to construct a lightweight model for image enhancement. However, those methods are limited by optimizing single MSE loss and it will cause some blurry and unrealistic enhancement results.

1.2 Deep residual learning

Deep learning firstly attracts great attention in Ref. [10], and it showed significant promotions in image classification tasks. Then, VGG networks were presented in Ref. [11], and they become universal feature extraction models. Ref. [12] proposed Inception network to introduce multi-scale feature representation in CNN. Ref. [12] demonstrated that deeper network can accordingly achieve better performance. Afterwards, many work focus on increasing depth of CNN to promote performance. However, when deeper networks are able to start converging, a degradation problem is exposed, that is, with the network depth increasing the performance gets saturated and then degrades rapidly. Besides, vanishing gradient problem still limits the performance of CNN.

Residual learning tries to solve those problems by constructing identical mapping, and the depth of CNN is substantially increased. It can be written as $\boldsymbol{y} = \boldsymbol{x} + F(\boldsymbol{x})$, where $\boldsymbol{x}$ and $\boldsymbol{y}$ are the input and output vectors of the layers, and F represents the residual mapping to be learned. The ideology of residual learning can be integrated into many previous networks[8], and many image-to-image translation tasks also adopt residual learning method to abridge the gap of generated images and input images. Ref. [8] proposed a residual learning based CNN for image denoising. In Ref. [6], a residual dense network was proposed for image super-resolution. However, residual learning has some bottlenecks in image enhancement. The identical mapping $\boldsymbol{x}$ cannot force $F(\boldsymbol{x})$ to learn detailed difference between low-quality and high-quality images. Therefore, non-identical mapping is considered to adjusts input $\boldsymbol{x}$ to an appropriate value.

1.3 Perceptual loss

A high-quality image should have clear textures, abundant colors, and conform to human perception. Thus, Ref. [13] introduced a pre-trained VGG network to compute perceptual loss for improving the quality of generated images. Ref. [9] proposed an enhancement method based on perceptual loss, which enriches more high-frequency information of enhanced images. Ref. [14] proposed generative adversarial nets (GANs), which has become an effective way for image generation. A conditional GAN was proposed in Ref. [15] for image-to-image translation task. Ref. [16] proposed a cycle-consistent adversarial networks for style transfer. Super-resolution based GAN adopts a generator, a feature extractor, and a discriminator to optimize hybrid loss, and they also achieve state-of-the-art performance in human perceptions. However, real-world image enhancement is a universal task for various image transformations (texture, luminance and resolutions). In Ref. [17], universal image enhancement frameworks were proposed. They publish a new large-scale image enhancement dataset based on DSLR camera. And a multi-term loss function is composed of color, texture and content terms, allowing an efficient image quality estimation. For image enhancement, optimizing high-level perceptual loss tends to extrude the shape of objects, while optimizing low-level perceptual loss can generate color-bright images. However, conventional multi-level perceptual loss lacks of flexibility in balancing those two aspects, because they allocate a fix loss weight for each level. Those weights are dynamically controlled to promote the flexibility.

2 The proposed method

2.1 General framework

The architecture of non-identical residual learning for image enhancement via dynamic multi-level perceptual loss is shown in Fig. 1. The holistic color adjustment globally adjusts the low-quality image to high-quality target. The fine-grained residual generation can recover texture and color details. Conventional residual learning[17] for image enhancement can be represented as

$$I^{H} = \tau I^{L} + G_{\theta}(I^{L}) \tag{1}$$

where $I^{L} \in R^{3\times H\times W}$ denotes the low-quality image and $I^{H} \in R^{3\times H\times W}$ denotes the enhanced high-quality image, G_{θ} is the proposed attention-aware recursive network to generate residual RGB image and τ is a constant. In the framework, the overall formulation of the non-identical residual learning for image enhancement can be written as

$$I^{H} = \boldsymbol{Y} \odot I^{L} + G_{\theta}(I^{L}) \tag{2}$$

where $\boldsymbol{Y} \in R^{3\times H\times W}$ is a trainable matrix rather than a single value. In the fine-grained residual generation, an attention-aware recursive network is proposed to generate fine residuals, and it is composed of three components. In the first component, the fully multi-scale block (FMSB) aims to extract multi-scale primary features. By N-step recursions in the recursive block, deep feature representations can be exploited. Finally, the reconstruction component converts the deep features to residual image. The generated residual image will be added with the adjusted image for getting the final enhanced image. In network training, three losses, i. e. MSE loss, dynamic multilevel perceptual loss (DPL) and adversarial loss are utilized.

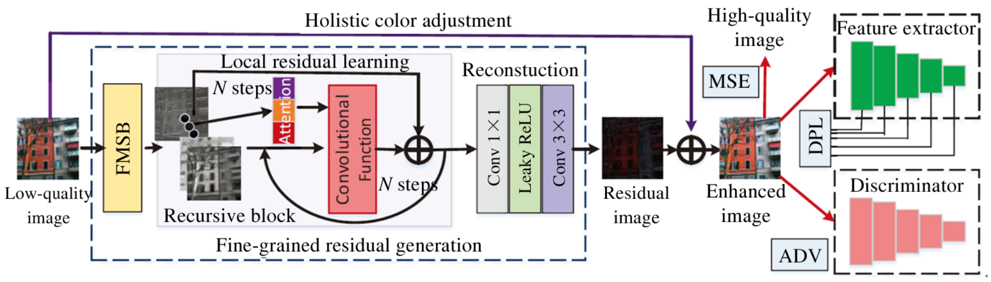


Fig. 1 Framework of the proposed method (the holistic color adjustment globally adjusts low-quality image, so the fine-grained residual generation tends to generate elaborate details. FMSB denotes fully multi-scale block, and ⊕ denotes element-wise addition. DPL denotes dynamic multi-level perceptual loss, and ADV is adversarial loss)

2.2 Holistic color adjustment

According to Eq. (1), conventional residual learning sets τ always is 1, which is called identical mapping. It cannot effectively force the attention-aware recursive network to learn elaborately detail differences, resulting in color deviations and texture losses in enhanced images. In the non-identical residual learning framework, an adjusting rate λ is firstly utilized to replace τ. According to Ref. [18] and the experiments, the negative residual is detrimental for network optimization. Thus, $\frac{1}{WHC}\sum_{m=1}^{H}\sum_{q=1}^{W}\sum_{p=1}^{C}(I^H_{p,q,m}-\lambda I^L_{p,q,m})$ is set to larger than 0 to generate positive residuals integrally from the attention-aware recursive network. On the other hand, the pixel-wise MSE loss is usually adopted in image enhancement. It can be formulated as

$$L_{\text{mse}} = \frac{1}{WHC}\sum_{m=1}^{H}\sum_{q=1}^{W}\sum_{p=1}^{C}\|(I^H_{p,q,m}-\lambda I^L_{p,q,m}) - G(I^L_{p,q,m})\|_2^2 \qquad (3)$$

Given an input low-quality image and its corresponding high-quality image, the $G(I^L_{p,q,m})$ are very small constant. Thus, $(I^H_{p,q,m}-\lambda I^L_{p,q,m})$ is adjusted to minimize L_{mse}. The computation of L_{mse} is not heavy, and the computational complexity is $O(n^3)$. An appropriate λ should be calculated to make this term equal to 0.

$$\frac{1}{WHC}\sum_{m=1}^{H}\sum_{q=1}^{W}\sum_{p=1}^{C}\|I^H_{p,q,m}-\lambda I^L_{p,q,m}\|_2^2 = 0 \qquad (4)$$

Conventional residual connections set λ is always 1, thus it cannot acquire a very optimal solution. The rational λ should be computed as

$$\lambda = \frac{1}{WHC}\sum_{m=1}^{H}\sum_{q=1}^{W}\sum_{p=1}^{C} I^H_{p,q,m}/I^L_{p,q,m} \qquad (5)$$

The result of λ is a fixed statistical value determined by training dataset as shown in Fig. 2. It should be computed by

$$\lambda = 1 + \frac{1}{n}\sum_{i=1}^{n}(Mean(P^L_{(i)}) - Mean(P^H_{(i)})) \qquad (6)$$

where $P^L_{(i)}$ and $P^H_{(i)}$ denote a low-quality image patch and a high-quality image patch; n is the total number of selected patches. However, in the simulations, this adjusting method treats every pixel equally and ignores the content of image. Consequently, some pixels can be inappropriately adjusted. λ should be viewed as a matrix rather than a fixed value, and the matrix should be adaptively produced according to input images. Inspired by the traditional image enhancement method[3], a low-quality image can be obtained by its high-quality version as

$$I^L = I^H \odot T \qquad (7)$$

where $T \in R^{3\times H\times W}$ represents an adjusting map in the method, and $\odot$ means element-wise multiplication. For each pixel x, the globally adjusted representation $I^{H-}(x)$ is calculated by

$$I^{H-}(x) = I^L(x)/(T(x) + \epsilon) \qquad (8)$$

where ϵ is a very small constant to avoid the zero denominator. $I^{H-}(x)$ is used instead of I^H, because $I^{H-}(x)$ is a coarse enhanced result, and it should be added with residual image for final high-quality generation.

$$I^H(x) = I^{H-}(x) + G_\theta(I^L(x)) \qquad (9)$$

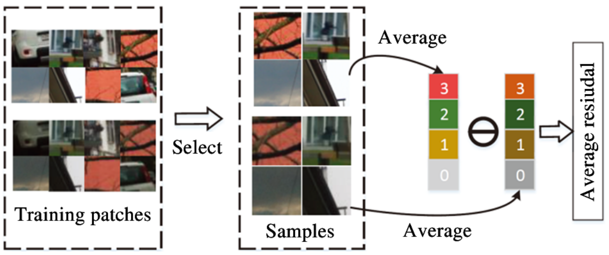


Fig. 2 Flow diagram of average residual computation

For obtaining T, a matrix $\boldsymbol{A} \in R^{3\times H\times W}$ is defined, in which all values are set to λ stated in Eq. (6). It provides a standard to prevent improper T. Then, features $X^{(0)} \in R^{64\times H\times W}$ of low-quality image is extracted as sources for adjusting map estimation. Let's define two matrixes are randomly initialized, i. e. $\boldsymbol{J}_1^{\mathrm{T}} \in R^{64\times 96\times 1\times 1}$

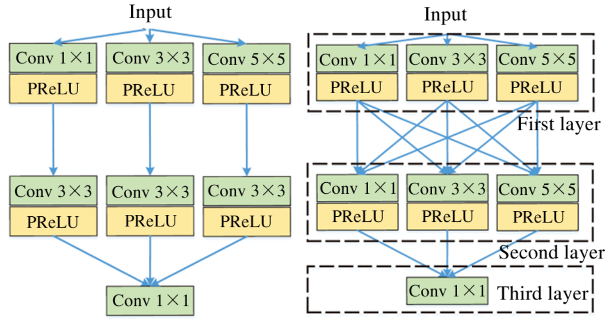
and $\boldsymbol{J}_2^{\mathrm{T}} \in R^{96\times3\times1\times1}$ to convert $X^{(0)}$ to a 3-channel output. T is accordingly computed as

$$T = \boldsymbol{A} + X^{(0)}\boldsymbol{J}_1^{\mathrm{T}}\boldsymbol{J}_2^{\mathrm{T}} \qquad (10)$$

$X^{(0)}$ can be obtained from FMSB and will be illustrated in the next section. The parameters of adjusting map T_θ can be updated with the network together.

2.3 Fine-grained residual generation

Fully multi-scale block. The success of inception network[4] has shown that the multi-scale information can provide multiple views to detect one image. So, the extracted features can benefit final image reconstruction. Motivated by Ref. [4], a fully multi-scale block is designed to extract primary features. It is composed of two multi-scale convolutional layers and a compressive layer, as shown in Fig. 3. In the first layer, convolutional kernels are adopted with three sizes $W_{i\times i}^{(1)}(i \in \{1,3,5\})$ to extract multi-scale features, in which PReLU is selected as activation functions[19]. Each convolutional kernel introduces all multi-scale features from the first layer, which utilizes features extracted by three kinds of receptive field. While the conventional multi-scale block only utilizes one kind of receptive field. Thus, FMSB can obtain more abundant information from the first layer to bring diversity representation. Finally, 1 × 1 convolutional kernel is used to compress feature maps and perform non-linear mapping.

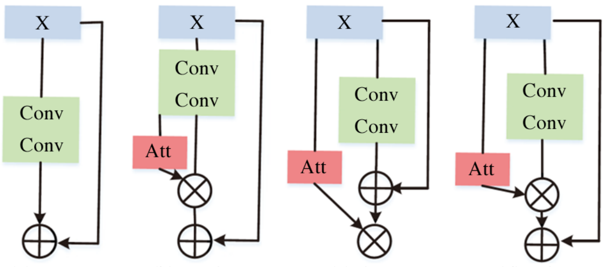


(a) Conventional multi-scale block (b) Proposed fully multi-scale block

Fig. 3 Comparison of the proposed fully multi-scaleblock against the conventional multi-scale block

Recursive block. Recently, residual recursive structures are proposed and show promising performance in super-resolution tasks[9]. It can construct large receptive fields by reusing convolutional layers. However, a well-behaved image enhancement model should flexibly consider different input conditions (light and color etc.), and feature representations in conventional recursive structures are limited due to parameters sharing. So some dynamic factors are introduced in this structure by adaptively selecting appropriate channels according to input images. Based on this motivation, attention mechanisms are employed[20] and three kinds of attention-aware recursive units are built.

Design of recursive units. A recursive block consists of multiple recursive units. Fig. 4(a) shows a typical recursive structure proposed in Ref. [8], which has no attention mechanism. In this work, three kinds of attention based recursive units are designed to explore the effectiveness of dynamic factors in different weighting scope. Their architectures are listed in Fig. 4(b), (c), and (d). Fig. 4(b) is residual recursive attentive unit (RRAU), which aims to effectively extract local discriminative features via directly weighting the convolutional features of the input image X. Fig. 4(c) is attentive residual recursive unit (ARRU), which aims to adaptively select global recursive features via weighting the residual recursive representations. Fig. 4(d) is residual attentive recursive unit (RARU), which aims to enhance the mutual information between convolutional features and inputs via second-order residual attentive weighting. According to the experiments, RARU is more appropriate for image enhancement task. Hence, RARU is used in block construction. Double 3 × 3 convolution and ReLU are stacked to construct a convolutional function for convolving.



(a) RRU (b) RRAU (c) ARRU (d) RARU

Fig. 4 Four types of recursive units

Attention along with recursion. To flexibly adapt to different image scenes, the feature representations in every recursive step should have domain consistency and be beneficial for the high-quality reconstruction. Inspired by Ref. [20], an attention along with recursive structures is designed to provide more flexible feature representations. Specifically, an adaptive average pooling function f_{AP} is firstly adopted to obtain global information for each feature map:

$$H^{(0)} = f_{AP}(X^{(0)}) \qquad (11)$$

where $H^{(0)} \in R^{1\times c}$ is the average result of $X^{(0)}$. Then, $H^{(0)}$ is embedded into low-dimension space:

$$E^{(i)} = Embed(H^{(0)}) = \sigma(\boldsymbol{W}_E^{(i)}H^{(0)}) \qquad (12)$$

where $\boldsymbol{W}_E^{(i)}$ denotes the parameter matrices of linear layer, and $\boldsymbol{W}_E^{(i)} \in R^{c \times c/2}$, σ is tangent function.

A reconstruction matrix $\boldsymbol{W}_R^{(i)} \in R^{c/2 \times c}$ is used to rebuild the original dimension vectors. The purpose is to encourage the attention module to learn meaningful weights from low dimension vectors. $\boldsymbol{A}$ is defined as reconstruction function. The output attention weights can be computed:

$$a^{(i)} = \boldsymbol{A}(E^{(i)}) = \text{softmax}(\boldsymbol{W}_R^{(i)} E^{(i)}) \qquad (13)$$

Finally, the attention weights perform an element-wise product with $X_{w/o}^{(i)}$:

$$X_{att}^{(i)} = X_{w/o}^{(i)} \cdot a^{(i)} \qquad (14)$$

The multi-level channel attention can constantly add attentive information by $a^{(i)}$ along with recursions.

Local residual learning. It is introduced to further improve information flow and transmit gradient for recursive structure. Therefore, element-wise addition is taken for $X_{att}^{(i)}$, and the output of the i step can be written as

$$X^{(i+1)} = X^{(0)} + X_{att}^{(i)} \qquad (15)$$

The local residual learning is designated to always start from $X^{(0)}$ for efficient and stable training[19]. Notably, due to the global non-identical residual learning can adjust the above-mentioned residual gap, local identical residual learning is normally adopted in the recursive block.

2.4 Dynamic multi-level perceptual loss

An individual MSE loss based optimization approaches usually lead to generating blurry and unrealistic results in image-to-image translation[16]. Inspired by Ref. [17], the generative adversarial nets (GANs) are considered. They are accomplished by utilizing an adversarial loss, which minimizes KL-divergence between the distribution of images produced by the generator and the distribution of images in the training dataset. An adversarial learning framework based on dynamic multi-level perceptual loss is proposed, which mainly contains a attention-aware recursive generator, a pre-trained VGG-19-based feature extractor, and a CNN-based discriminator. Specially, the feature extractor and the discriminator are used as two constraints to optimize the enhanced images generated by the generator from low-quality images. Among them, the feature extractor provides dynamic multi-level perceptual loss of hierarchical content, and the discriminator provides the measure of similarity between the generated images and corresponding ground-truths.

The feature extractor can provide perceptual loss based on content error between enhanced images and their high-quality versions. However, conventional GAN based methods for image enhancement usually optimize high-level perceptual loss, losing accuracy in color representations. Based on the motivation that optimizing high-level perceptual loss is beneficial for recovering spatial and texture information, and optimizing low-level perceptual loss is helpful for color reconstruction[9], hierarchical features are utilized, which are widely applied to classification[21] and detection[22] tasks. Instead of solely optimizing high-level content loss, five content losses are optimized from the output of each max-pooling layer cooperatively. In this way, the generated patches tend to be more consistent with human perception. The overall formulation of the multi-level perceptual loss is

$$L_p = \sum_{i=1}^{I} a_i L_c^{(i)} \qquad (16)$$

where, a_i is a weight for the ith level and $L_p^{(i)}$ denotes the corresponding perceptual loss.

However, the weight a_i is usually hard to design due to uncertainty of the importance of each level perceptual loss. Although equally allocating weights is an intuitive way, but the importance of each level perceptual loss is also dynamic with training process. Hence, a weight a_i can be computed via a dynamic way based on error feedback ideology. Firstly, the ith average content error z_i between generated patches P^G and high-quality patches P^H is computed. Then the weight a_i is gotten via a softmax function for normalizing those errors z_i:

$$z_i = \frac{1}{WHC} \sum_{m=1}^{H} \sum_{q=1}^{W} \sum_{p=1}^{C} \| \phi_i(P^G)_{m,q,p} - \phi_i(P^H)_{m,q,p} \|_1 \qquad (17)$$

$$a^{(i)} = \frac{e^{z_i}}{\sum_{i=1}^{5} e^{z_i}} \qquad (18)$$

where ϕ_i is a non-linear function for dynamically computing the content error between P^G and P^H. H, W, and C are the feature size. e denotes nature exponential. Notably, the structure of the discriminator is referenced in Ref. [22].

3 Experiments

3.1 Dataset and metrics

Following Ref. [17], the classic DSLR enhancement dataset (DPED) is adopted to train and test the method. The DSLR is specially collected for image enhancement tasks. The image quadruples in DSLR are captured by cameras with different qualities. Detailedly, DPED contains 4549 photos from Sony smartphone, 5727 photos from iPhone, and 6015 photos from Canon. The peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) index are two prevailing criteria selected for evaluations. In the experiments, PSNR and SSIM are all calculated on RGB space. In ablation studies, the most challenging iPhone dataset[25] is selected to conduct validation and 400 pairs of patches for testing are randomly selected. Without loss of generality, NRL is optimized by MSE loss. The detailed settings are introduced in next section.

3.2 Ablation study for network architectures

Fully multi-scale block (FMSB) can better extract features compared with conventional multi-scale block (MSB). Other three methods are compared: (1) double convolutional layer with 3 × 3 kernel size (Conv-3); (2) a large convolutional layer with 9 × 9 kernel size (Conv9); (3) multi-scale block (MSB) as shown in Fig. 4; (4) fully multi-scale block, in which the second convolutional layers are all 3 × 3 (FMSB-3). In line with the settings in Ref. [12], ResNet is used as backbone network. As shown in Table 1, FMSB achieves the highest scores both in PSNR and SSIM and it is 0.04 dB PSNR and 0.0025 SSIM higher than MSB.

Table 1 Results of different blocks on the iPhone dataset

Method	Conv3	Conv9	MSB	FMSB-3	FMSB
PSNR	22.35	22.31	22.37	22.38	22.41
SSIM	0.8868	0.8897	0.8876	0.8862	0.8901

It demonstrates the superiority of FMSB. Conv3 and Conv9, large kernel size tends to achieve higher SSIM but low PSNR. FMSB-3 gains a very close PSNR to FMSB, but FMSB has different sizes of kernel in the second layer, which can better exploit holistic color information. In summary, the proposed FMSB is an effective block for primary feature extraction.

To verify the effectiveness of the attention along with recursive architectures, a comparison for different designs is shown in Fig. 4. Besides, RRU + A is introduced which adds single attention mechanism in the last recursive step. Without loss of generality, PSNR is compared for verifying modeling capability in 1, 3, 6, 12, 24 recursive steps (*N* steps). Experimental results are listed in Table 2.

Table 2 PSNR results of different recursive steps

Method	RRU	RRAU	ARRU	RRU + A	RARU
$N = 1$	22.45	22.75	22.69	-	22.72
$N = 3$	22.53	22.81	22.74	22.77	22.78
$N = 6$	22.50	22.81	22.67	22.81	22.86
$N = 12$	22.56	22.73	22.72	22.78	22.82
$N = 24$	22.60	22.79	22.73	22.82	22.91

It can be seen that except for $N = 1$ and $N = 3$, RARU achieves the best performance compared with the others. From $N = 6$ to $N = 12$, it can be seen that simply increasing recursive steps sometimes can decrease performance. Especially, RRAU degrades performance with *N* increasing. It can be seen that the setting of 6-step recursions is cost-effective. The 6-step RARU achieves promising performance meanwhile has less recursive steps. It outperforms RRU, RRAU and ARRU by 0.36, 0.05, and 0.05 dB PSNR, respectively. Training curves are exhibited in the condition of 6-step recursions in Fig. 5. The pink curve is conventional ResNet structure as illustrated in Ref. [17]. It can get similar results before 20 epochs. While RARU can stably increase PSNR, resulting in the best performance. Therefore, RARU is considered as the final structure.

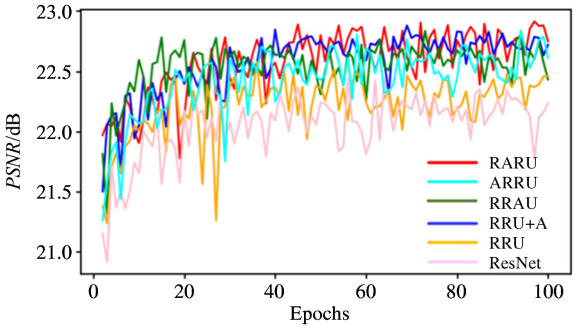


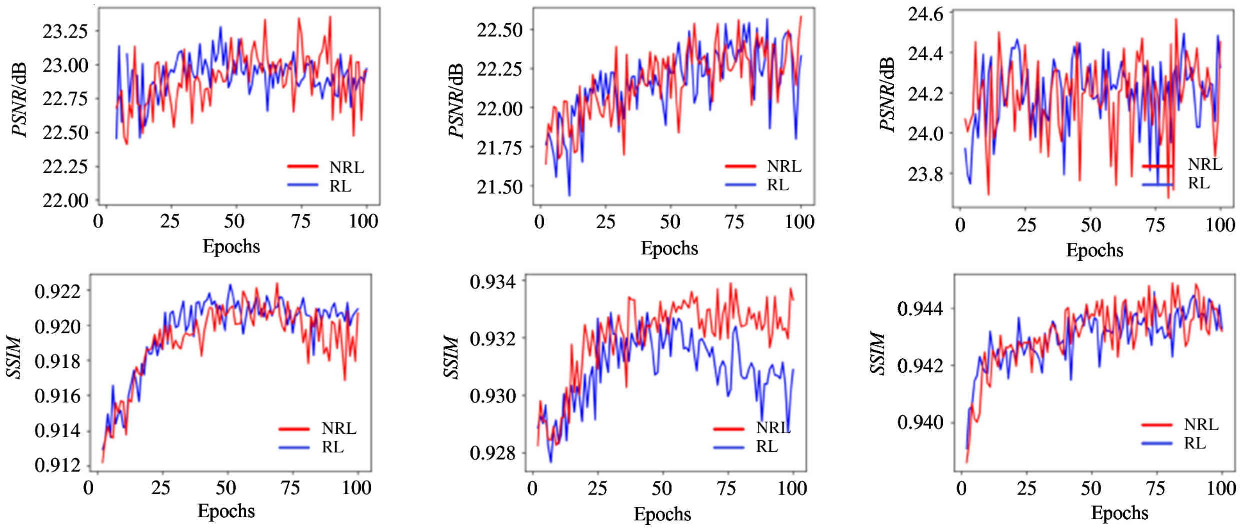
Fig. 5 PSNR testing results with different recursive structures

3.3 Evaluation for non-identical residual learning

For evaluating the non-identical residual learning (denoted as RL), three control group are set, i. e. residual learning is not utilized in the holistic color adjustment (denote as no-RL), conventional residual learning, incomplete non-identical residual learning as illustrated in Eq. (6) (denoted as NRL-). Experimental results are shown in Table 3. Compared with no-RL and RL, RL outperforms no-RL by 0.47 dB PSNR and 0.0097 SSIM on the iPhone dataset. The advantage of residual learning is clearly demonstrated. The performance of NRL- is slightly lower than RL, because some pixels cannot be accurately adjusted. While the NRL achieves the best performance, it outperforms RL by 0.09 dB PSNR and 0.013 SSIM on the Sony dataset, respectively. Although both RL and NRL achieve identical PSNR value (22.54 dB) on the BlackBerry dataset, NRL precedes RL by 0.0027 SSIM, showing the effectiveness of NRL. The training curve of RL and NRL are visualized in Fig. 6. It can be seen that the NRL lines are higher than the RL lines in the most conditions. Though both RL and NRL cause unstable PSNR curves on the Sony dataset, the NRL is still higher than RL in some peak values. Conclusively, the non-identical residual learning is an effective method, which is superior to conventional residual learning in the image enhancement task.

Table 3 Experimental results of residual learning

Method	no-RL (PSNR/SSIM)	RL (PSNR/SSIM)	NRL- (PSNR/SSIM)	NRL (PSNR/SSIM)
iPhone	22.86/0.9114	23.33/0.9211	23.26/0.9190	23.36/0.9211
BlackBerry	22.38/0.9245	22.54/0.9306	22.44/0.9277	22.54/0.9333
Sony	24.20/0.9401	24.48/0.9435	24.42/0.9421	24.57/0.9448



(a) Results conducted on the iPhone dataset (b) Results conducted on the BlackBerry dataset (c) Results performed on the Sony dataset

Fig. 6 Comparisons of non-identical residual learning and conventional residual learning with $N = 6$

3.4 Comparisons with state-of-the-art methods

The proposed method is compared with the state-of-the-art methods (Apple photo enhancer (APE) is taken as a baseline). Ref. [23] was a 3-layer CNN and was optimized by MSE. Ref. [9] was classical image-to-image translation method based on perceptual losses. Refs[17,24] were all state-of-the-art enhancement methods. Ref. [25] was an adversarial learning framework and its generator is replaced by the attention-aware recursive network for fair comparison. NRL denotes the non-identical residual learning framework optimized by individual MSE loss. Besides, NRL is introduced with the proposed dynamic multi-level perceptual loss (denotes as NRL-DPL). Experimental results are shown in Table 4, where NRL-DPL achieves the highest SSIM among all others. Concretely, NRL-DPL outperforms by 0.0072[17] and 0.0046[24] SSIM on the iPhone dataset, respectively. It also outperforms by 3.82 dB[23] and 1.14 dB[25] PSNR on the Sony dataset, respectively. It demonstrates the state-of-the-art performance of the method for image enhancement. NRL also achieves favourable PSNR results compared with others. It reveals the strong generalization ability of the network architecture. NRL-DPL outperforms NRL except for PNSR on the BlackBerry dataset, showing superiority of the proposed DPL and adversarial learning strategy. According to Fig. 7, the method achieves better visual effect and less unpleasant artifacts. In the first group comparison, the bag can be enhanced more distinctly by the method.

Table 4 Comparisons with the state-of-the-art methods in PSNR/SSIM

Method	APE (baseline)	Ref. [21]	Ref. [9]	Ref. [17]	Ref. [24]	Ref. [23]	NRL	NRL-DPL
iPhone	17.28/0.8631	19.27/0.8992	20.32/0.9161	21.35/0.9201	22.69/0.9205	22.52/0.9227	23.34/0.9210	23.38/0.9273
BlackBerry	18.91/0.8922	18.89/0.9134	20.11/0.9298	20.66/0.9328	21.97/0.9331	22.39/0.9336	22.53/0.9333	22.48/0.9354
Sony	19.45/0.9168	21.21/0.9382	21.33/0.9434	22.01/0.9437	23.89/0.9428	23.86/0.9461	24.55/0.9448	25.03/0.9477

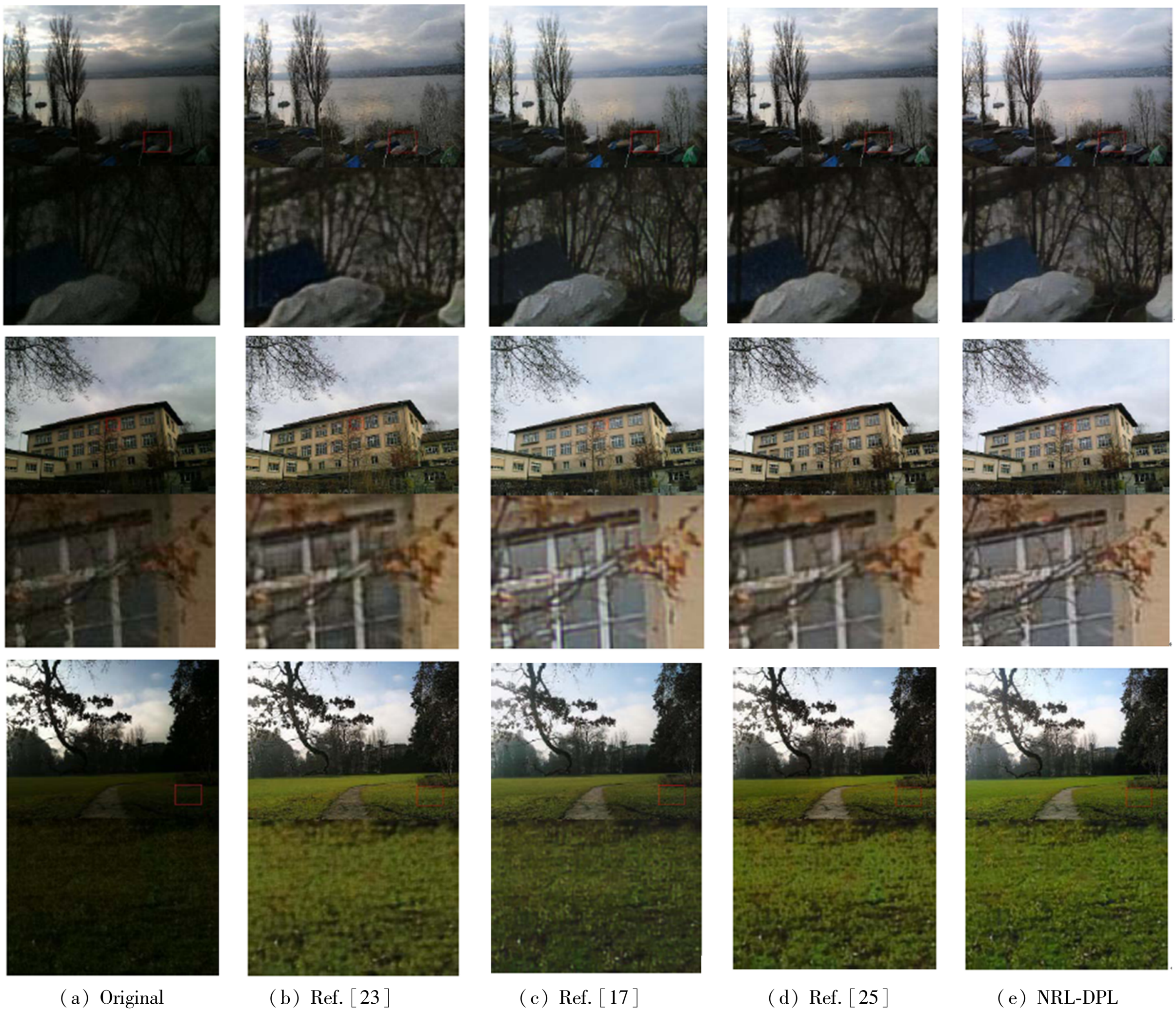


Fig. 7 Examples of visual enhancement comparisons on DPED

In the second group, the edges of the window has less artifacts compared with Ref. [17]. In the last group, the method achieves a very high-quality result both in overall and local details. Notably, the model has only 190×10^3 parameters compared with 400×10^3 in Ref. [17]; it also demonstrates the advantage of the proposed recursive architecture.

The proposed framework is trained with 6 recursive steps without batch normalization (BN). All channel numbers are set to 64 in the recursive block. 1/6 patches are randomly selected in training dataset as one epoch. The Adam is adopted for optimizing the network, and the initial learning rate is set to 0.0005. Training batch is set to 32. For each 5 epochs, the learning rate will decrease by the scale of 0.95. Training is stopped at 100 epoch. Experiments are performed on double NVIDIA Titan XP GPUs for training and testing. The training process costs about 14 h for 100 epochs, and the average testing speed of a 256 × 256 patch is 0.04 s.

3.5 User study

Previous classical work are followed to perform mean opinion score (MOS) tests, which quantify the ability of different approaches to re-construct perceptually convincing images. 100 low-quality images are selected from VOC2012 (VOC2012-LQ100) for testing. Specifically, 29 raters are asked for assigning an integral score from 1 (bad quality) to 5 (excellent quality). The score criterion is four-fold: Color(Col), Texture (Tex), Luminance (Lumin), Overall (Over). Four methods are evaluated, i. e., Ref. [17], Ref. [23], Ref. [25] and the method NRL-DPL. They are all trained on the iPhone dataset for evaluating their adaptability. According to results in Table 5, the proposed method achieves the highest average MOS scores. Although Ref. [17] achieved the highest luminance score, it causes many harsh textures. Overall, the method performs the best scores in color, texture and overall feeling.

Fig. 8 shows some visual examples. Apparently, the enhanced images obtained by the method have more perceptual comfortableness and textural softness.

Table 5 MOS testing results on VOC2012-LQ100

Method	Col	Tex	Lumin	Over	Average
Original	2.0	2.7	1.6	2.9	2.30
Ref. [23]	3.1	2.6	2.0	2.5	2.55
Ref. [17]	3.0	1.8	3.8	3.1	2.93
Ref. [25]	3.3	2.7	3.3	3.4	3.18
NRL-DPL	3.9	3.0	3.4	3.6	3.48

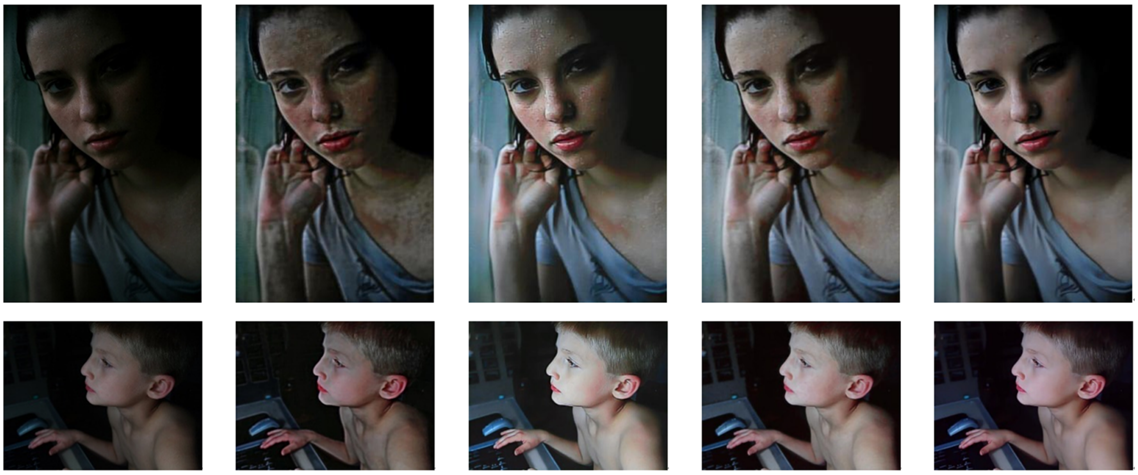

(a) Original (b) Ref.[23] (c) Ref.[17] (d) Ref.[25] NRL-DPL

Fig. 8 The selected visual demonstration on VOC2012-LQ100

4 Conclusions

In this paper, a non-identical residual learning for image enhancement via dynamic multi-level perceptual loss is proposed, which views image enhancement as two branches. In the first branch, a holistic color adjustment method is designed to adjust global color representation to the high-qualities. It forces the second branch to accurately capture color and texture details by learning elaborate difference. In the second branch, an attention-aware recursive network is proposed to adaptively transform features according to image color conditions, as well as mitigate overfitting problem. Last but not least, a dynamic multi-level content loss is designed to improve color effect as high-quality images. Extensive experiments conducted on publicly available datasets demonstrate the state-of-the-art performance of the proposed method.

References

[1] DEVIKA G, PARTHASARATHY S. Optimised transformation model for contrast enhancement of satellite images using modified whale optimisation[J]. *International Journal of Image and Data Fusion*, 2018, 10(2): 131-145

[2] HARDIE R C, EISMANN W T, WILSON G L. MAP estimation for hyperspectral image resolution enhancement using an auxiliary sensor[J]. *IEEE Transactions on Image Processing*, 2004, 13(9):1174-1184

[3] GUO X, LI Y, LING H. LIME: low-light image enhancement via illumination map estimation[J]. *IEEE Transactions on Image Processing*, 2017, 26(2): 982-993

[4] LORE K G, AKINTAYO A, SARKAR S. Llnet: a deep autoencoder approach to natural low-light image enhancement[J]. *Pattern Recognition*, 2017, 61(1):650-662

[5] YU J, GAO X, TAO D, et al. A unified learning framework for single image super-resolution[J]. *IEEE Transactions Neural Networks and Learning System*, 2014, 25(4):780-792

[6] HUANG T. Neighbor2Neighbor: self-supervised denoising from single noisy images[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Los Angeles, USA, 2021:14781-14790

[7] JEON J, LEE S. Reconstruction-based pairwise depth dataset for depth image enhancement using CNN[C]//European Conference on Computer Vision, Munich, Germany, 2018: 438-454

[8] HAN D, KIM J. Deep pyramidal residual networks[C]//IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA, 2017: 6307-6315

[9] JOHNSON J, ALAHI A, LI F F. Perceptual losses for real-time style transfer and super-resolution[C]//European Conference on Computer Vision, Amsterdam, Netherlands, 2016:694-711

[10] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks[C]// Annual Conference on Neural Information Processing Systems, Lake Tahoe, USA, 2012:1106-1114

[11] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[C]// International Conference on Learning Representations, San Diego, USA, 2015: 1-10

[12] SZEGEDY C, LIU W, JIA Y, et al. Going deeper with convolutions[C]//IEEE Conference on Computer Vision and Pattern Recognition, Boston, USA, 2015:1-9

[13] ANCUTI C O, ANCUTI C A, VLEESCHOUWER C D, et al. Color balance and fusion for underwater image enhancement[J]. *IEEE Transactions on Image Processing*, 2018, 27(1):379-393

[14] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[J]. *Neural Information Processing Systems*, 2014, 45(3): 2672-2680

[15] ISOLA P, ZHU J Y, ZHOU T, et al. Image-to-image translation with conditional adversarial networks[C]// IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA, 2017:1125-1134

[16] ZHU J Y, PARK T, ISOLA P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks[C]// International Conference on Computer Vision, Venice, Italy, 2017: 2223-2232

[17] IGNATOV A, KOBYSHEV N, TIMOFTE R, et al. Dslr-quality photos on mobile devices with deep convolutional networks[C]//IEEE International Conference on Computer Vision, Venice, Italy, 2017:3277- 3285

[18] HE K, ZHANG X, REN S, et al. Identity mappings in deep residual networks[C]// European Conference on Computer Vision, Amsterdam, Netherlands, 2016:630-645

[19] LEDIG C. Delving deep into rectifiers: surpassing human-level performance on imagenet classification[C]// IEEE International Conference on Computer Vision, Santiago, Chile, 2015:1026-1034

[20] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]//IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018:7132-7141

[21] TANG P, WANG X, SHI B, et al. Deep fishernet for image classification[J]. *IEEE Transactions on Neural Networks and Learning System*, 2019, 30(7):2244-2250

[22] ZHAO Z Q, ZHENG P, XU S T, et al. Object detection with deep learning: a review[J]. *IEEE Transactions on Neural Networks and Learning System*, 2019, 13(4): 112-119

[23] DONG C, LOY C C, HE C, et al. Image super-resolution using deep convolutional networks[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016, 38(2):295-307

[24] LIU J, JUNG C. Multiple connected residual network for image enhancement on smartphones[C]//European Conference on Computer Vision Workshops, Munich, Germany, 2018:182-196

[25] ZHU X, LI Z, ZHANG X, et al. Generative adversarial image super-resolution through deep dense skip connections[J]. *Computing Graph Forum*, 2018,37(7): 289-300

HU Ruiguang, born in 1984. He is currently a senior engineer in Beijing Aerospace Automatic Control Institute. He received his Ph. D degree from Institute of Automation, Chinese Academy of Science in 2014. He also received his M. S. and B. S. degrees from Chongqing University in 2009 and 2006, respectively. His research interests include machine learning, image processing, object recognition, and intelligent control.