Applied Intelligence (2022) 52:652–661
https://doi.org/10.1007/s10489-021-02489-x

Efficient residual attention network for single image super-resolution

Fangwei Hao[1] · Taiping Zhang[1] · Linchang Zhao[1] · Yuanyan Tang[2]

Accepted: 29 April 2021 / Published online: 8 May 2021

Abstract
The use of deep convolutional neural networks (CNNs) for image super-resolution (SR) from low-resolution (LR) input has achieved remarkable reconstruction performance with the utilization of residual structures and visual attention mechanisms. However, existing single image super-resolution (SISR) methods with deeper network architectures can encounter representational bottlenecks in CNN-based networks and neglect model efficiency in model statistical inference. To solve these issues, in this paper, we design a channel hourglass residual structure (CHRS) and explore an efficient channel attention (ECA) mechanism to extract more representative features and ease the computational burden. Specifically, our CHRS, consisting of several nested residual modules, is developed to learn more discriminative representations with fewer model parameters, and the ECA is presented to efficiently capture local cross-channel interaction by subtly applying 1D convolution. Finally, we propose an efficient residual attention network (ERAN), which not only fully learns more representative features but also pays special attention to network learning efficiency. Extensive experiments demonstrate that our ERAN achieves certain improvements in model performance and implementation efficiency compared to other previous state-of-the-art methods.

Keywords Image super-resolution · Channel hourglass residual structure · Efficient channel attention mechanism · Efficient residual attention network

1 Introduction

Because of the strong representation abilities of deep convolutional neural networks (CNNs), deep CNN-based networks, including pioneering residual network [1], feature pyramid network [2] and stacked hourglass network [3], have achieved great progress in computer vision tasks such as object classification [1, 4, 5], target detection [6–9] and many other endeavors [3, 10–14]. In recent years, single image super-resolution (SISR) [15], which aims to recover visual high-resolution (HR) output from low-resolution (LR) input, has drawn much attention from researchers. While there always exists an ill-posed problem where the same LR image can be downsampled from diverse HR images, many significant CNN-based networks [17–23] have emerged in SISR for modeling the nonlinear mapping function from an LR image to HR more accurately. Dong et al. [16] first designed a three-layer CNN named SRCNN to model the nonlinear mapping function and obtained surprising performance. For further improvements of reconstruction, Kim et al. [17] designed a deeper network whose depth reached 20 and achieved high effectiveness. After the appearance of the pioneering residual network [1], Lim et al. [18] modified the general residual module and proposed a more complex network termed EDSR, which obtained notable performance but encountered many model parameters. Then, the dense SR model RDN [19], which utilized hierarchical features by dense connection, was presented, but its performance was similar to EDSR. Later, more advanced networks were built, including RCAN [20] and SAN [21], which both introduced an attention mechanism into SR models. Although they obtained a significant learning capacity for a CNN by stacking modified residual modules and introducing the general channel attention (CA) mechanism to learn the interdependencies among feature channels, they seldom focused

✉ Fangwei Hao
fangweihao@cqu.edu.cn

Taiping Zhang
tpzhang@cqu.edu.cn

Linchang Zhao
anilue@cqu.edu.cn

Yuanyan Tang
yytang@umac.mo

[1] College of Computer Science, Chongqing University, Chongqing 400044, China

[2] Faculty of Science and Technology, University of Macau, Zhuhai, China

on learning discriminative representations with a more efficient residual module and rarely considered modeling channel-wise interactions efficiently. Recently, Lan et al. [22] proposed a network with a dual global pathway named ERN, the designed local wider residual block in which the batch normalization (BN) layers were removed expanded wider channels before the activation layer; as a result, the expanded wider channels increased the number of parameters. These deep networks cannot learn discriminative features while maintaining fewer model parameters; that is, they are not efficient.

To address these limitations, we propose an efficient residual attention network (ERAN) to improve the model's learning effectiveness and efficiency. We propose a channel hourglass residual structure (CHRS) to deepen the residual block and generate a nested residual block for extracting discriminative features efficiently. To the best of our knowledge, our CHRS is the first to apply the hourglass structure among feature channels. Furthermore, we present an efficient channel attention (ECA) mechanism to model the channel-wise interdependencies of features. Then, we integrate this mechanism into our CHRS and generate an efficient residual attention block (ERAB). Finally, we use a Laplacian pyramid framework similar to [23] to build our SR network.

In summary, there are three contributions offered in this work:

- We propose an efficient residual attention network (ERAN) to reconstruct high-performance HR image from the corresponding LR. Our ERAN is much deeper than most previous CNN-based networks and achieves better SR performance while reducing model parameters to some extent.
- We propose a channel hourglass residual structure (CHRS) to deepen the residual block and generate nested residuals for accelerating information flow, bypassing massive low-frequency information and learning discriminative representation efficiently.
- We propose an efficient channel attention (ECA) mechanism to drive the model to efficiently learn the channel-wise interdependencies in the SISR network.

The remainder of this paper is organized as follows: the next section presents an overview of the related work. Section 3 describes the proposed model in detail. Section 4 shows the empirical research results. Section 5 presents the conclusion.

2 Related work

In recent years, unprecedented progress has been made in deep image super-resolution. The pioneering CNN-based SR work proposed by Dong et al. [16] employed a three-layer CNN to learn the mapping function from LR images to HR images and was termed SRCNN. Benefiting from the prediction performance of the CNN, its results showed great improvements when quantitatively and visually compared with the early interpolation-based method [24]. To increase the learning capacity of the network, Kim et al. [17] deepened the depth of the network to 20 and obtained remarkable SR performance. As skip connections were proposed in CNN networks [1, 25], much deeper models rapidly emerged. Lim et al. [18] designed a very wide and deep network named EDSR by stacking many modified residual blocks. Their network achieved significant improvements in performance and demonstrated the significance of model depth in image SISR. Other deep SR works, such as RDN [19] and SRDenseNet [26], which were derived from the dense-connection network [25], paid more attention to utilizing hierarchical features from different convolution layers. Their operations, stemming from densely concatenating features of different layers, increased the reuse of features and enabled further feature fusion. To achieve better visual SR performance, Ledig et al. [27] proposed SRGAN, which was based on a generative adversarial network (GAN) [28] and combined perceptual and adversarial loss with l2 loss. Although the blurring and oversmoothing artifacts were alleviated to a certain extent by applying SRGAN, its reconstruction results may not have been faithful because of the produced unpleasing artifacts. Then, Lan et al. [21] expanded wider channels in general residual block removed batch normalization (BN) layers and proposed one deep network with a dual global pathway named ERN.

An attention mechanism can generally be regarded as allocating available processing resources towards the most informative part of input. Massive works integrated with attention mechanisms have been proposed for different tasks, including image classification [29] and SISR [20, 21]. To resolve the limitation of network depth and explore the general channel attention (CA) mechanism in SISR, Zhang et al. [20] designed a very deep RCAN network composed of many residual channel attention blocks (RCABs) and residual in residual (RIR) structures. An RIR structure can drive the model to bypass abundant low-frequency information and reconstruct more accurate results. SAN [21] introduced a second-order channel-wise attention module and a nonlocal attention mechanism and combined them with an effective residual structure; eventually, the network successfully captured discriminative representations and long-distance spatial contextual information. Although both methods obtain notable improvements quantitatively and visually when integrated with the general CA mechanism, they are burdened with heavy computational costs.

Recently, Wang et al. [30] proposed an efficient channel attention (ECA) block in the classification task to efficiently model channel-wise interdependencies across feature maps and obtained accurate performance with fewer parameters.

However, there are few proposed works that explore the impact of ECA on SISR.

3 Our model

To make full use of the powerful representation of the residual module and efficient channel-wise mechanism in the SISR task, we design a deep advanced residual network integrated with the ECA mechanism and name it an efficient residual attention network (ERAN) (see Fig. 1).

3.1 Network architecture

As shown in Fig. 1, our ERAN is mainly made up of four parts: shallow feature extraction, efficient residual blocks (ERABs) for deep feature extraction, upscale modules of SR levels and corresponding reconstruction blocks. Let us suppose that I_{LR} and I_{SR} represent the input and output of our network, respectively. Similar to [18, 20, 27], given I_{LR} as the input, we extract its shallow feature maps F_0 using only one convolutional layer (Conv)

$$F_0 = H_f(I_{LR}) , \tag{1}$$

where $H_f(\cdot)$ is the convolution operation.

Similar to [23], our model consists of $B = log_2(S)$ reconstruction levels, where S denotes the scale factor, i.e., the ×2 network has 1 level, and the ×4 network has 2 levels and so on. There are M ERABs at each level in our network. The first ERAB at level b extracts features from its input, and the extracted features act as the input of the next ERAB at the same level. The output of the last ERAB at level b denotes acquired abstract features at the current level, so we altogether have B groups of abstract features from corresponding B levels

$$F_{DF-b} = H_{ERAB-M}\left(H_{ERAB-(M-1)}\left(\cdots H_{ERAB-1}\left(F_{up-(b-1)}\right)\right)\right) , \tag{2}$$

where F_{DF-b}, H_{ERB-M} and $F_{up-(b-1)}$ represent the acquired abstract features at level b, the M-th ERAB operation at level b and the upsampled feature maps at level $b-1$, respectively. Then, the deep abstract features F_{DF-b} are upscaled by the upscale module at the b level

$$F_{up-b} = H_{up-b\uparrow}(F_{DF-b}) , \tag{3}$$

where $H_{up-b\uparrow}$ and F_{up-b} are the upscale module and upscaled feature maps at level b, respectively. There are several choices for upscaling models, such as transposed convolution [31] and ESPCN [32], in which good trade-offs between computation and performance are obtained by applying these post-upscaling strategies. Following [20, 21], we adopt sub-pixel convolution [32] in our upscale model. Next, we use one convolution layer at each level to reconstruct the result at the current level.

There are some available choices for the loss function to optimize the SR model, such as L1 [18, 20–22], L2 [16, 17], perceptual and adversarial losses [27]. For fair comparisons with advanced methods [20–22], we also choose the L1 loss function for model optimization. Hence, the objective function of ERAN is defined as:

$$L(\Theta) = \sum_{b=1}^{B} \frac{1}{N} \sum_{i=1}^{N} \left\| H_{ERAN-b}\left(I^i_{LR-b}\right) - I^i_{HR-b} \right\|_1 , \tag{4}$$

where Θ is the parameter set of our model. For fast and effective convergence in the training process, the Adam optimization algorithm [33] is adopted to optimize the complex network.

3.2 Channel hourglass residual structure (CHRS)

The hourglass network [3] is a novel design with the ability to capture diverse feature maps and fuse them together. It can generate pixel-wise predictions, which coincide with the goal

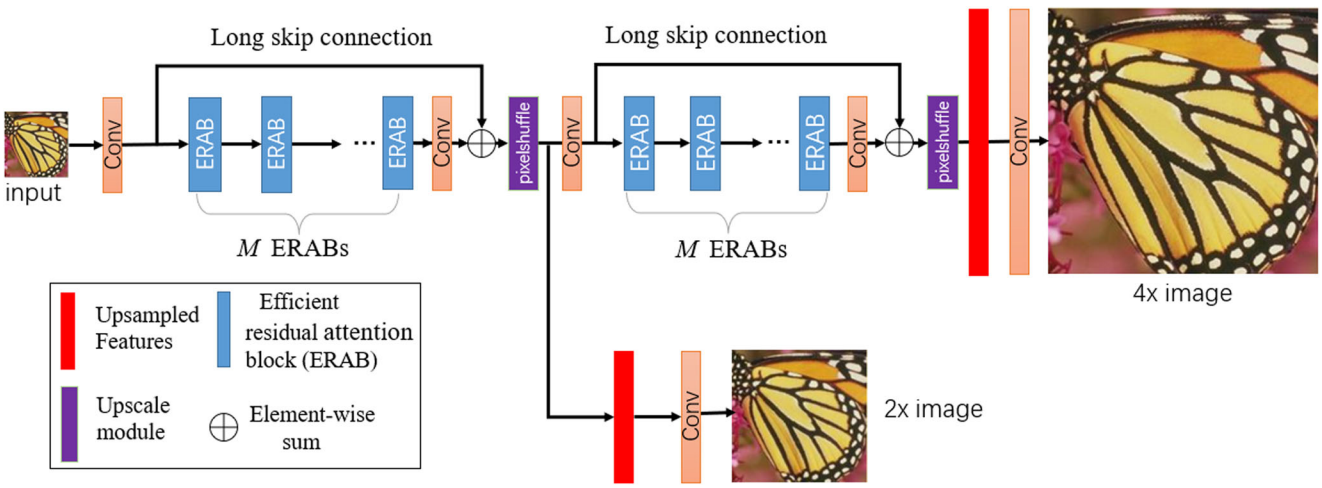

Fig. 1 Network architecture of our ERAN for 4× SR.

of the SISR task. Motivated by the theory [1, 3, 17, 19] that a deeper network can obtain a more abstract expression and a residual in residual (RIR) structure can accelerate information flow and bypass abundant low-frequency information in the LR inputs, we subtly design a deeper channel hourglass residual structure, i.e., the CHRS (see Fig. 2), which consists of P nested residuals for image SR.

We now show more details about our CHRS. Suppose F_{input} denotes input feature maps with C channels and $H \times W$ size. The channels of the later layer in CHRS are halved to $C/2$ while keeping the $H \times W$ size unchanged at all times. After intermediate feature maps reach the fewest channels, i.e., $\frac{C}{2^P}$, the CHRS starts twofold increasing convolution kernels to double the channels and combines corresponding cross-scale feature maps by P element-wise additions. These RIR operations can make the CHRS bypass abundant low-frequency information and capture powerfully expressive information. Table 1 clearly shows the difference in efficiency between the general residual module [1] removed BN layers and our CHRS. Our CHRS has fewer parameters but a larger module depth and more residual connections under the same input size and output size. Note that the feature resolutions of different layers in our CHRS are all the same, which makes the CHRS be easily extended to other state-of-the-art SR networks. These dense residual connections across different layers accelerate the information flow and make the CHRS focus on high-frequency information during model training. Different from the usage of ReLU in [20], in our CHRS, all convolution layers except the last are followed by the LeakyReLU activation function.

3.3 Efficient Channel attention (ECA) module

In this section, we revisit the general channel attention (CA) mechanism and clarify more details about the ECA module (see Fig. 3).

3.3.1 Revisiting Channel attention (CA) mechanism

Suppose that given feature maps $\mathbf{X} = [x_1, x_2, \cdots, x_c]$ with C channels and $H \times W$ size, global average pooling is used to learn the channel-wise global statistic information $\boldsymbol{z}$. Then, we can obtain the c-th value of $\boldsymbol{z}$ by

$$z_c(x_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} x_c(i,j) \ , \tag{5}$$

where $x_c(i,j)$ denotes the pixel value of the c-th feature map x_c at spatial position (i,j). Then, a sigmoid gating mechanism is adopted in [20, 21] to capture the channel-wise weights

$$\widehat{z} = \sigma(W_U \delta(W_D z)) \ , \tag{6}$$

where $\sigma(\cdot)$ and $\delta(\cdot)$ denote the sigmoid gating function and ReLU function, respectively, and W_U and W_D are the weight settings of the channel-upscaling layer and the channel-downscaling layer, respectively. To avoid high computing complexity, W_D are often set to $C \times \left(\frac{C}{r}\right)$, and W_U are set to $\left(\frac{C}{r}\right) \times C$. Although convolution operations that change the numbers of convolution kernels limit model complexity in the CA module, the channel information and its weight are not directly corresponding.

3.3.2 Efficient channel attention (ECA) mechanism

The ECA mechanism (see Fig. 3) is motivated by the general channel attention (CA) mechanism used in the RCAN; it models interdependencies among feature channels adaptively and efficiently by considering local cross-channel interaction. The ECA module investigates one 1D convolution layer with an adaptive kernel size to replace the two 2D convolution layers in the general CA module and makes the network focus on capturing powerful feature maps efficiently.

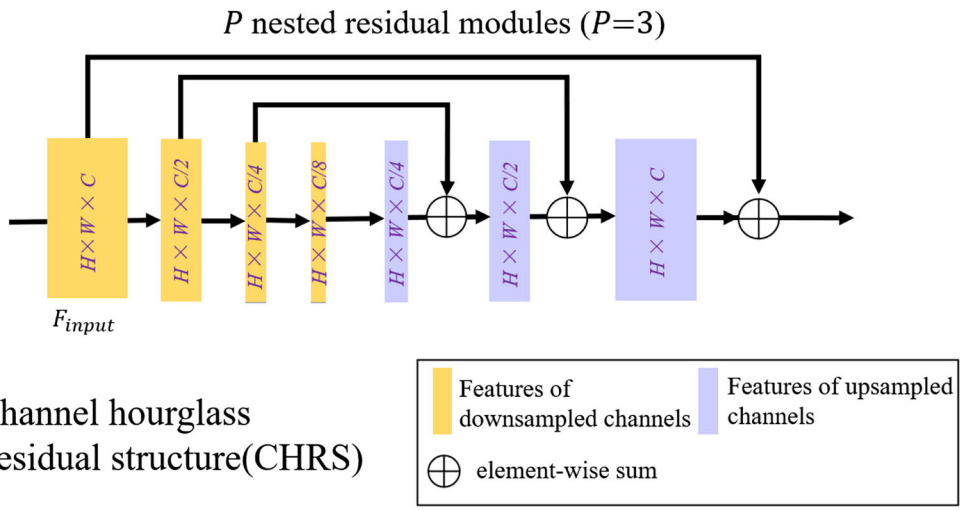

Fig. 2 The architecture of our channel hourglass residual structure (CHRS), consisting of P=3 nested residuals, makes the depth of CHRS reach 6

Table 1 Efficiency comparison between general residual module removed BN layers and our CHRS

Module	Kernel size	Input size	Output size	#.Param	depth
general res. [1]	k=3	C×H×W	C×H×W	2*k*k*C*C=A	2
CHRS(P=2)		C×H×W	C×H×W	5/4*k*k*C*C=0.625A	4
CHRS(P=3)		C×H×W	C×H×W	21/16*k*k*C*C=0.656A	6

Given the feature maps $\boldsymbol{z} \in R^C$ without reducing the dimension, channel-wise weights can be obtained by

$$\boldsymbol{\alpha} = \sigma(\boldsymbol{W} \times \boldsymbol{z}) , \tag{7}$$

where $\boldsymbol{W}$ and $\sigma(\cdot)$ are parameter matrices with the dimension of $C \times C$ and a sigmoid gating function, respectively. To capture the discriminative representation among feature channels efficiently, the key step is how to model the local cross-channel interaction. Considering z_i and its k neighbors, the weight of z_i can be calculated by

$$\alpha_i = \sigma\left(\sum_{j=1}^{k} w^j z_i^j\right), z_i^j \in \Omega_i^k , \tag{8}$$

where Ω_i^k is the group of k adjacent channels of z_i. In brief, such local aggregation can be exactly implemented by 1D convolution with a kernel size of k

$$\boldsymbol{\alpha} = \sigma(conv_{1D}(\boldsymbol{z})) , \tag{9}$$

where $conv_{1D}(\cdot)$ is a 1D convolution layer and its kernel size equals k.

Hence, the remaining key issue is how to set the value of k. Considering the similar philosophy, feature maps with different channel dimension C should reasonably have different statistical values of k; therefore, a mapping function $\phi(\cdot)$ may be available from k to C

$$C = \phi(k) , \tag{10}$$

Generally, a linear function, i.e., $\phi(k) = \gamma * k - q$, is usually adopted to model the simplest corresponding mapping. However, the simple linear function limits the expression of complicated relations between k and C. To better describe the complex quantitative relations, we introduce a nonlinear function, i.e.,

$$C = \phi(k) = 2^{(\gamma * k - q)} , \tag{11}$$

to replace the linear one. The reason why an exponential function is used is that the channel dimension C of feature maps is usually set to a power of 2. Then, given a channel dimension value of C, the kernel size k can be calculated adaptively by

$$k = \varphi(C) = \left| \frac{log_2(C)}{\gamma} + \frac{q}{\gamma} \right|_{odd} , \tag{12}$$

where $|t|_{odd}$ is the odd number nearest to t. Following [30], in our experiments, γ and q are always set to 2 and 1, respectively. Clearly, using nonlinear mapping $\varphi(\cdot)$ gives feature maps with different channel numbers different range interactions

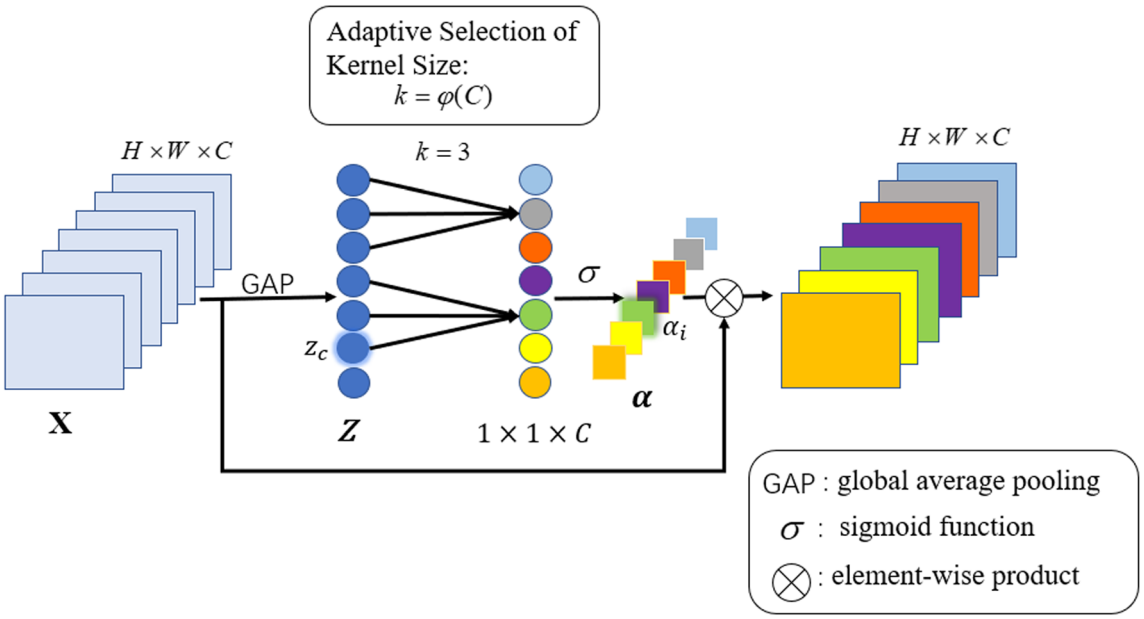

Fig. 3 Efficient channel attention (ECA) module used in our ERAN

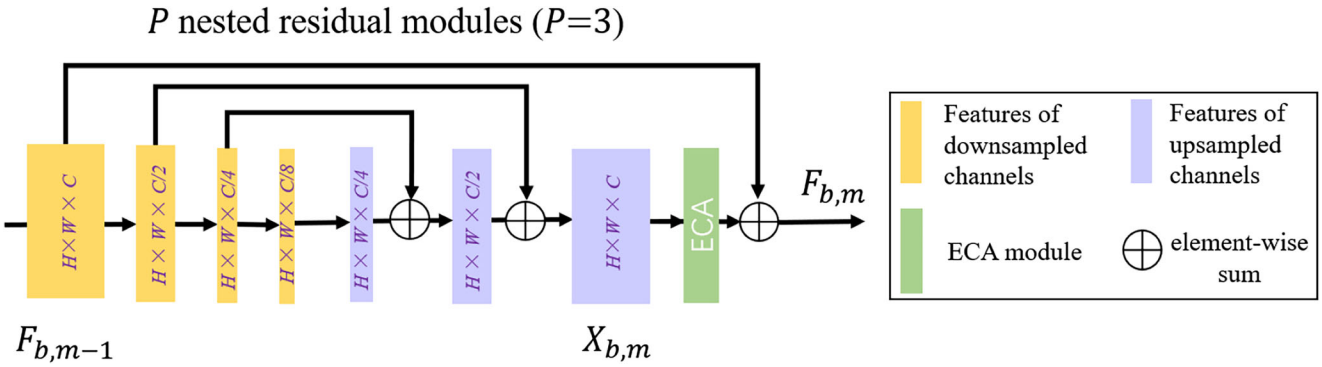

Fig. 4 The architecture of the proposed efficient residual block (ERAB)

and drives the model to adaptively learn the interdependencies among feature channels.

3.4 Efficient residual attention block (ERAB)

To take advantage of the feature maps with channel-wise weights effectively, we incorporate the ECA mechanism into our CHRS and generate an efficient residual attention block (ERAB) (see Fig. 4) to learn discriminative representation.

Inspired by the effectiveness of residual blocks and residual in residual (RIR) structure in [20], long skip connections are added into our model to enhance information flow in the network. For the m-th ERAB at the b-th level, we have

$$\begin{cases} F_{b,m} = F_{b,m-1} + R_{b,m}\left(X_{b,m}\right) \\ R_{b,m}\left(X_{b,m}\right) = \sigma_{b,m}\left(\mathrm{conv}_{1D}^{b,m}\left(GAP_{b,m}\left(X_{b,m}\right)\right)\right)\cdot X_{b,m} \end{cases}, \tag{13}$$

where $R_{b,\,m}(\cdot)$ indicates the function of efficient channel attention (ECA), and its components $GAP_{b,\,m}(\cdot)$, $conv_{1D}^{b,m}(\cdot)$ and $\sigma_{b,\,m}(\cdot)$ are the global average pooling function, 1D convolution layer and corresponding sigmoid gating function, respectively. $F_{b,\,m-1}$ and $F_{b,\,m}$ denote the input and output of the m-th ERAB in which the residual $X_{b,\,m}$ is learned after the input feature maps $F_{b,\,m-1}$ are dealt with by $P-1$ residual subunits. Considering the trade-off between the performance of our ERAB and module computation, in our experiments, P is always set to 3.

3.5 Joint optimization with added losses

Our network architecture with multiple SR levels is similar to the Laplacian pyramid framework [23], but we use our ERABs to extract deep features. In addition, we only obtain the SR result from the last level, i.e., the results of internal levels are only used to supervise and optimize the result at the last level. Theoretically the same LR image can be downsampled from infinite HR images, and there are many possible functions to choose in mapping function space. To alleviate the learning diversity for the deep model, we adopt a network architecture similar to the Laplacian pyramid framework so that internal levels can help the model learn the mapping function from LR to HR image more accurately.

At each SR level of our model, there are M ERABs and one sub-pixel convolution layer. Each sub-pixel convolution layer is connected to a corresponding convolution layer to recover the HR image at the current level. For ×4 and ×8 SR models, M is always set to 30.

4 Experimental results

In this section, we first clarify our experimental settings in detail, including datasets, evaluation metrics, optimizer and related equipment. Then, we verify the contribution of each component and the impact from different combinations of components in the proposed ERAN. We show the results quantitatively and visually compared with other advanced methods. Finally, we present a model complexity analysis, including the parameters of different models.

4.1 Settings

Following [34], we train our networks on DIV2K [35] and Flickr2K [18] datasets. After training, we test our models on

Table 2 Effects of CHRS and ECA; the best PSNR (dB) values on Set5 (4×) are observed in 1×10^4 iterations

Channel attention (CA)	×	√	×	×	√	×
Efficient Channel Attention (ECA)	×	×	√	×	×	√
Channel hourglass residual structure (CHRS)	×	×	×	√	√	√
PSNR on Set5 (4×)	32.59	32.62	32.63	32.60	32.64	32.66

Table 3 Quantitative results with BI degradation model. The best and second-best results are highlighted and underlined, respectively

Algorithms	Scale	Set5 PNSR/SSIM	Set14 PNSR/SSIM	BSD100 PNSR/SSIM	Urban100 PNSR/SSIM	Manga109 PNSR/SSIM
Bicubic	×4	28.42/0.8104	26.00/0.7027	25.96/0.6675	23.14/0.6577	24.89/0.7866
SRCNN [16]		30.48/0.8628	27.50/0.7513	26.90/0.7101	24.52/0.7221	27.58/0.8555
VDSR [17]		31.35/0.8830	28.02/0.7680	27.29/0.0726	25.18/0.7540	28.83/0.8870
LapSRN [23]		31.54/0.8850	28.19/0.7720	27.32/0.7270	25.21/0.7560	29.09/0.8900
SRDenseNet [26]		32.02/0.8941	28.50/0.7783	27.54/0.7332	26.05/0.7813	29.50/0.8992
EDSR [18]		32.46/0.8968	28.80/0.7876	27.71/0.7420	26.64/0.8033	31.02/0.9148
ERN [22]		32.39/0.8975	28.75/0.7853	27.70/0.7398	26.43/0.7966	–/–
RDN [19]		32.47/0.8990	28.81/0.7871	27.72/0.7419	26.61/0.8028	31.00/0.9151
RCAN [20]		32.63/0.9002	28.87/0.7889	27.77/0.7436	26.82/0.8087	31.22/0.9173
SAN [21]		32.64/0.9003	28.92/0.7888	27.78/0.7436	26.79/0.8068	31.18/0.9169
ERAN (ours)		32.66/0.8999	28.92/0.7891	27.79/0.7429	26.86/0.8073	31.39/0.9172
ERAN+ (ours)		**32.71/0.9008**	**28.95/0.7897**	**27.83/0.7443**	**27.05/0.8124**	**31.62/0.9199**
Bicubic	×8	24.40/0.6580	23.10/0.5660	23.67/0.5480	20.74/0.5160	21.47/0.6500
SRCNN [16]		25.33/0.6900	23.76/0.5910	24.13/0.5660	21.29/0.5440	22.46/0.6950
VDSR [17]		25.93/0.7240	24.26/0.6140	24.49/0.5830	21.70/0.5710	23.16/0.7250
LapSRN [23]		26.15/0.7380	24.35/0.6200	24.54/0.5860	21.81/0.5810	23.39/0.7350
SRDenseNet [26]		25.99/0.7041	24.23/0.5810	24.46/0.5302	21.67/0.5619	23.10/0.7121
EDSR [18]		26.96/0.7762	24.91/0.6420	24.81/0.5985	22.51/0.6221	24.69/0.7481
RDN [19]		27.23/0.7854	25.25/0.6505	24.91/0.6032	22.83/0.6374	25.14/0.7994
RCAN [20]		27.31/0.7878	25.23/0.6511	24.98/0.6058	23.00/0.6452	25.24/0.8029
SAN [21]		27.22/0.7829	25.14/0.6476	24.88/0.6011	22.70/0.6312	24.85/0.7910
ERAN (ours)		27.32/0.7885	25.24/0.6497	24.96/0.6033	22.95/0.6391	25.24/0.8013
ERAN+ (ours)		**27.35/0.7896**	**25.27/0.6513**	**25.00/0.6045**	**23.11/0.6452**	**25.42/0.8049**

five benchmark datasets, including SET5 [36], SET14 [37], BSDS100 [38], URBAN100 [39] and MANGA109 [40], and adopt the peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) [41] on the Y channel as evaluation metrics after transforming the SR results to YCbCr space. We carry out extensive experiments with a bicubic (BI) degradation model and use scaling factors ×4 and ×8 for training and testing.

During training, the ADAM [33] optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.99$, and $\varepsilon = 10^{-8}$ is practically adopted to optimize our model. We conduct all experiments using Pytorch [42] on a computer equipped with one GTX 1080Ti GPU, one Intel i7-8700k CPU and 24 GB system memory. The learning rate is initially set to 10^{-4} and decays with a cosine annealing strategy.

4.2 Ablation investigation

We analyze the effects of the channel hourglass residual structure (CHRS) and efficient channel attention (ECA) mechanism compared with the channel attention (CA) mechanism and conduct a series of experiments to demonstrate the effectiveness of our network.

First, we train our model without the CHRS, ECA and CA on the DIV2K and Flickr2K datasets, and we obtain a basic performance value of 32.59 dB PNSR with general residual modules removed BN layers. Next, we carry out verification experiments with the CA, ECA or CHRS to analyze the effects and obtain corresponding results of 32.62 dB PNSR, 32.63 dB PNSR, and 32.60 dB PNSR, respectively. These clear results demonstrate the ability of each block to improve the model reconstruction performance. Then, we implement different experiments with different combinations of CA, ECA and CHRS. We observe that the model with the CA and CHRS achieves a 32.64 dB PSNR, which is better than the 32.62 dB PSNR of the module with CA only. The model with ECA and CHRS achieves a 32.66 dB PSNR, which is the best of these results. These findings show a powerful representation of our ERAB and the notable performance of our ERAN. All results are shown in Table 2.

4.3 Comparisons with advanced methods

To further verify the effectiveness of our ERAN, we conduct a large number of experiments and compare our results quantitatively and visually with other state-of-the-art methods, such as SRCNN [16], VDSR [17], LapSRN [23], EDSR [18], RDN

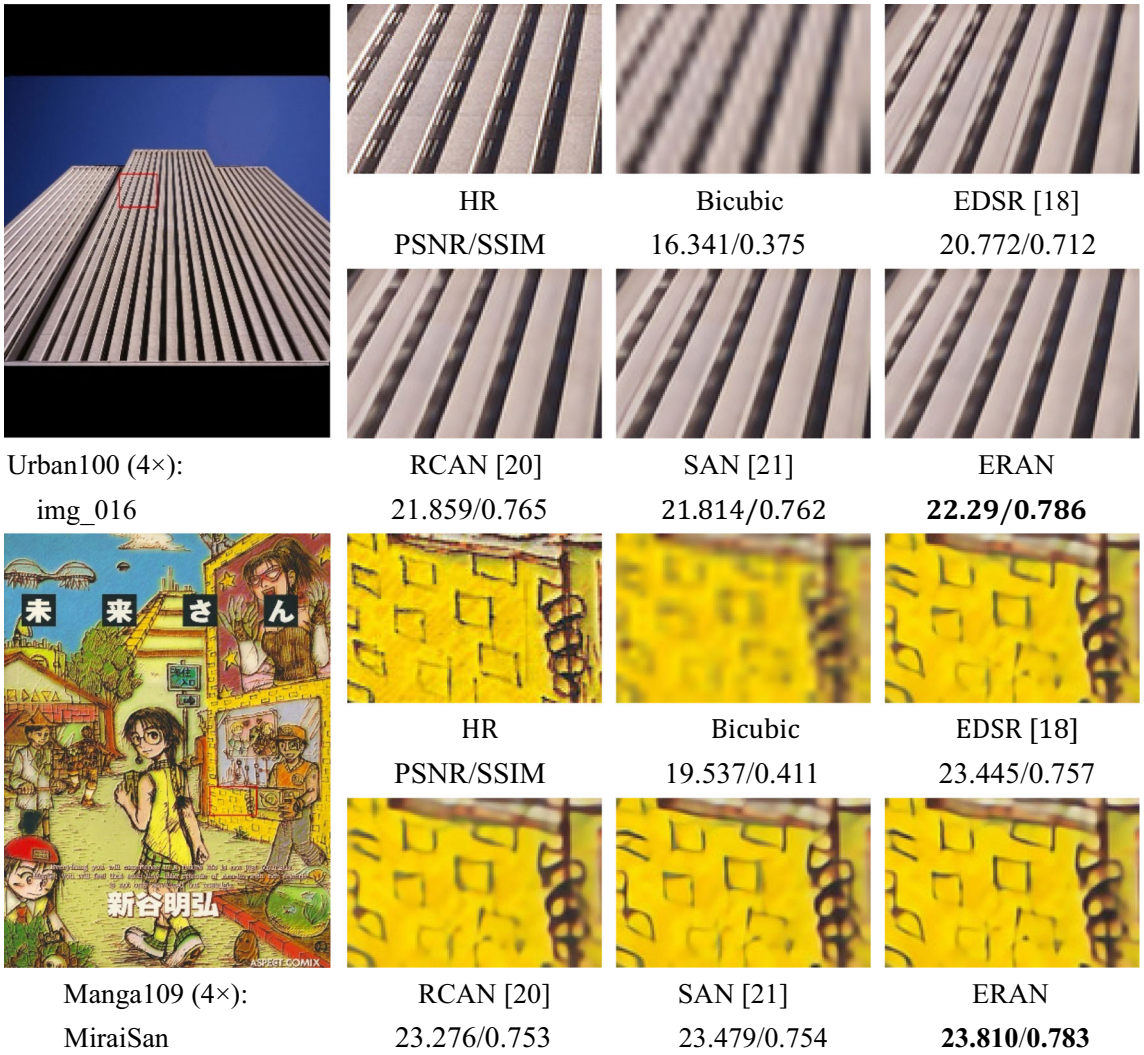

Fig. 5 Visual comparisons for 4× SR with the BI model on the Urban100 and Manga109 datasets. The best results are **highlighted**

[19], SRDenseNet [26], RCAN [20], SAN [21], and ERN [22]. Similar to [20, 21], the self-ensemble strategy is adopted to further improve our ERAN, denoted as ERAN+.

PSNR/SSIM results Quantitative evaluation results of ×4 and ×8 SR are shown in Table 3. For ×4 SR, our ERAN+ provides the best quantitative performance, with the highest PNSR and SSIM values on all datasets compared with previous advanced networks. Even without the self-ensemble strategy, our ERAN can yield comparable or superior results on five test datasets. In terms of a larger scaling factor (e.g., 8), our ERAN+ still achieves the best value of evaluation metrics, surpassing the outputs of the recent advanced CNN-based method SAN. All experimental records show that our model yields better performance than most state-of-the-art methods.

Visual results Figure 5 presents visual comparisons of SR scale ×4 on the datasets of Urban100 and Manga109. For image "img_016" and image "MiraiSan", the early bicubic method yields widespread blurring and even loses the main outlines. Other recent methods (e.g., EDSR, RCAN and SAN) can recover the main structure but have difficulty reconstructing clearer details and present some blurring artifacts or distorted edges. For our ERAN, it can be observed that

Table 4 Computation and parameter comparison (4× Set5)

Metric	LapSRN	EDSR	RDN	RCAN	ERN	SAN	RCAN+ERAB	ERAN (ours)
Paras (M)	0.813	43.1	22.3	15.6	9.53	15.9	13.31	8.02
PSNR(dB)	31.54	32.48	32.47	32.63	32.39	32.64	32.65	**32.66**

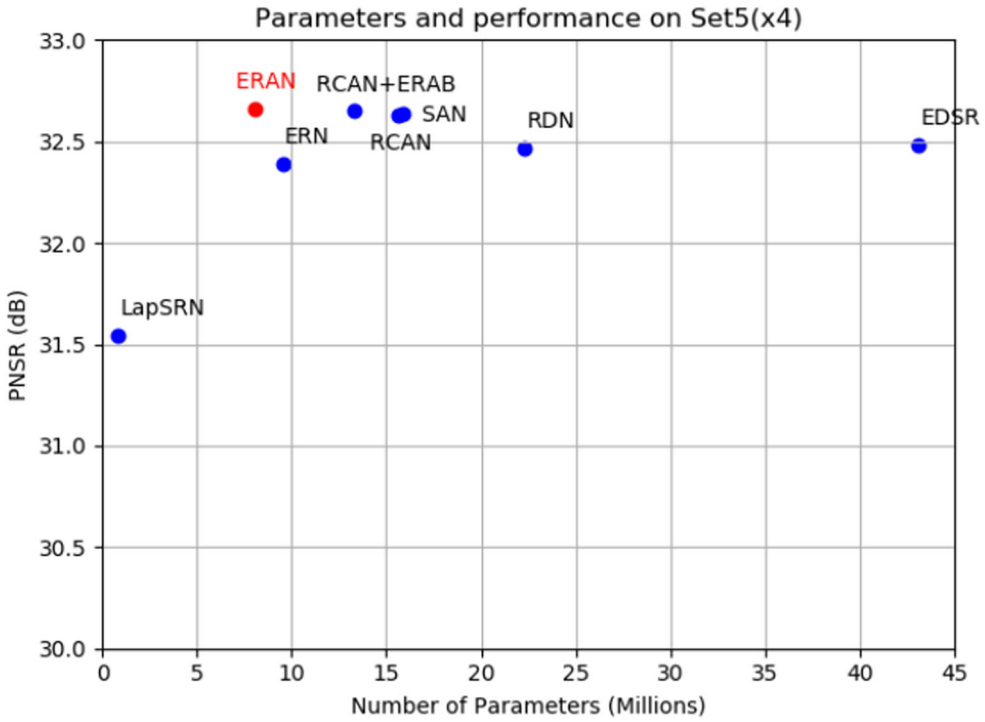

Fig. 6 Performance and the number of parameters on Set5

our model can recover more details, especially yield sharper edges, and more natural performance benefited from the better captured high-frequency information.

4.4 Model complexity analysis

Our goal is to obtain good performance with fewer parameters. The details of different advanced methods are shown in Table 4, and a corresponding visual illustration is presented in Fig. 6. We replace the residual channel attention block (RCAB) in RCAN with our ERAB, and the new RCAN model is denoted as RCAN+ERAB. RCAN+ERAB can obtain better performance with fewer parameters than RCAN for 4× SR on the Set5 dataset. In addition, our ERAN, with the fewest parameters, performs better than other state-of-the-art methods. This demonstrates the good trade-off of our ERAN between superior performance and model complexity.

5 Conclusions

We propose a very deep efficient residual attention network (ERAN) for accurate and efficient image SR. Specifically, the channel hourglass residual structure (CHRS) allows the ERAN to deepen the network by applying several nested residual modules, accelerate information flow and bypass massive low-frequency information from LR images by residual in residual (RIR) structure. In addition to designing the CHRS to learn discriminative representation with fewer model parameters, we propose an efficient channel attention (ECA) mechanism to efficiently learn channel-wise interdependencies by applying 1D convolution, and integrate this mechanism into the CHRS to generate an efficient residual attention block (ERAB). Extensive experiments on SISR with BI models demonstrate the effectiveness, efficiency of our ERAN and the generalization ability of our ERAB.

Acknowledgments The authors acknowledge the anonymous reviewers for their helpful comments.

References

1. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
2. Lin T Y, Dollár P, Girshick R, He K, Hariharan B, Belongie S (2017) Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2117–2125
3. Newell A, Yang K, Deng J (2016) Stacked hourglass networks for human pose estimation. In European conference on computer vision. Springer, Cham, pp 483–499
4. Cen F, Zhao X, Li W, Wang G (2021) Deep feature augmentation for occluded image classification. Pattern Recogn 111:107737
5. Qi C, Zhang J, Jia H, Mao Q, Wang L, Song H (2021) Deep face clustering using residual graph convolutional network. Knowl-Based Syst 211:106561
6. Tian Z, Shen C, Chen H, He T (2020) Fcos: a simple and strong anchor-free object detector. IEEE Transactions on Pattern Analysis and Machine Intelligence
7. Liu Y, Wang Y, Wang S, Liang T, Zhao Q, Tang Z, Ling H (2020) Cbnet: a novel composite backbone network architecture for object

detection. In: Proceedings of the AAAI conference on artificial intelligence, vol. 34, no. 07, pp 11653–11660
8. Li X, Song D, Dong Y (2020) Hierarchical feature fusion network for salient object detection. IEEE Trans Image Process 29:9165–9175
9. Li Z, Xi T, Zhang G, Liu J, He R (2021) AutoDet: pyramid network architecture search for object detection. Int J Comput Vis:1–19
10. Li X, Zhao H, Han L, Tong Y, Tan S, Yang K (2020) Gated fully fusion for semantic segmentation. In: Proceedings of the AAAI conference on artificial intelligence, vol. 34, no. 07, pp 11418–11425
11. Zhang H, Tian Y, Wang K, Zhang W, Wang FY (2019) Mask SSD: an effective single-stage approach to object instance segmentation. IEEE Trans Image Process 29:2078–2093
12. Quan Y, Chen Y, Shao Y, Teng H, Xu Y, Ji H (2021) Image denoising using complex-valued deep CNN. Pattern Recogn 111: 107639
13. Xu W, Song H, Zhang K, Liu Q, Liu J (2020) Learning lightweight multi-scale feedback residual network for single image super-resolution. Comput Vis Image Underst 197:103005
14. Koller O, Camgoz NC, Ney H, Bowden R (2019) Weakly supervised learning with multi-stream CNN-LSTM-HMMs to discover sequential parallelism in sign language videos. IEEE Trans Pattern Anal Mach Intell 42(9):2306–2320
15. Freeman WT, Pasztor EC, Carmichael OT (2000) Learning low-level vision. Int J Comput Vis 40(1):25–47
16. Dong C, Loy CC, He K, Tang X (2015) Image super-resolution using deep convolutional networks. IEEE Trans Pattern Anal Mach Intell 38(2):295–307
17. Kim J, Kwon Lee J, Mu Lee K (2016) Accurate image super-resolution using very deep convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1646–1654
18. Lim B, Son S, Kim H, Nah S, Mu Lee K (2017) Enhanced deep residual networks for single image super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp 136–144
19. Zhang Y, Tian Y, Kong Y, Zhong B, Fu Y (2020) Residual dense network for image restoration. IEEE Transactions on Pattern Analysis andMachine Intelligence
20. Zhang Y, Li K, Li K, Wang L, Zhong B, Fu Y (2018) Image super-resolution using very deep residual channel attention networks. In: Proceedings of the European conference on computer vision (ECCV), pp 286–301
21. Dai T, Cai J, Zhang Y, Xia S T, Zhang L (2019) Second-order attention network for single image super-resolution. In: proceedings of the IEEE conference on computer vision and pattern recognition, pp 11065–11074
22. Lan R, Sun L, Liu Z, Lu H, Su Z, Pang C, Luo X (2020) Cascading and enhanced residual networks for accurate single-image super-resolution. IEEE transactions on cybernetics
23. Lai W S, Huang J B, Ahuja N, Yang M H (2017) Deep laplacian pyramid networks for fast and accurate super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 624–632
24. Zhang L, Wu X (2006) An edge-guided image interpolation algorithm via directional filtering and data fusion. IEEE Trans Image Process 15(8):2226–2238
25. Huang G, Liu Z, Van Der Maaten L, Weinberger K Q (2017) Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4700–4708
26. Tong T, Li G, Liu X, Gao Q (2017) Image super-resolution using dense skip connections. In: Proceedings of the IEEE international conference on computer vision, pp 4799–4807
27. Ledig C, Theis L, Huszár F, Caballero J, Cunningham A, Acosta A, ..., Shi W (2017) Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4681–4690
28. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S et al (2014) Generative adversarial nets. Adv Neural Inf Proces Syst 27:2672–2680
29. Hu J, Shen L, Sun G (2018) Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7132–7141
30. Wang Q, Wu B, Zhu P, Li P, Zuo W, Hu Q (2020) ECA-net: efficient channel attention for deep convolutional neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 11534–11542
31. Dong C, Loy CC, Tang X (2016) Accelerating the super-resolution convolutional neural network. In European conference on computer vision (pp. 391-407). Springer, Cham
32. Shi W, Caballero J, Huszár F, Totz J, Aitken AP, Bishop R, ..., Wang Z (2016) Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1874–1883
33. Kingma D, Ba J. (2014) Adam: a method for stochastic optimization. Computer Science
34. Wang X, Yu K, Wu S, Gu J, Liu Y, Dong C, ..., Change Loy C (2018) Esrgan: Enhanced super-resolution generative adversarial networks. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 0–0
35. Timofte R, Agustsson E, Van Gool L, Yang M H, Zhang L (2017) Ntire 2017 challenge on single image super-resolution: methods and results. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp 114–125
36. Bevilacqua M, Roumy A, Guillemot C, Alberi-Morel ML (2012) Low-complexity single-image super-resolution based on nonnegative neighbor embedding
37. Zeyde R, Elad M, Protter M (2010) On single image scale-up using sparse-representations. In international conference on curves and surfaces (pp. 711-730). Springer, Berlin, Heidelberg
38. Arbelaez P, Maire M, Fowlkes C, Malik J (2010) Contour detection and hierarchical image segmentation. IEEE Trans Pattern Anal Mach Intell 33(5):898–916
39. Huang JB, Singh A, Ahuja N (2015) Single image super-resolution from transformed self-exemplars. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5197–5206
40. Matsui Y, Ito K, Aramaki Y, Fujimoto A, Ogawa T, Yamasaki T, Aizawa K (2017) Sketch-based manga retrieval using manga109 dataset. Multimed Tools Appl 76(20):21811–21838
41. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP (2004) Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process 13(4):600–612
42. Paszke A, Gross S, Chintala S, Chanan G, Yang E, DeVito Z, ..., Lerer A (2017) Automatic differentiation in pytorch