Expectation-Maximization Attention Cross Residual Network for Single Image Super-resolution

Xiaobiao Du[1]*, Jie Niu[2], Chongjin Liu[1]
[1]Zhuhai College of Jilin University
[2]Unit 61212 of the People's Liberation Army
xbiaodu@163.com, jieniuchina@163.com, chongjinliu@163.com

Abstract

The depth of deep convolution neural network and self-attention mechanism is widely used for the single image super-resolution (SISR) task. Nevertheless, we observed that the deeper network was more hard to train and the self-attention mechanism is computationally consuming. Residual learning has been widely recognized as a common approach to improve network performance for deep learning, but most existing methods did not make the best of the learning ability of deep CNN, thus hindering the ability of representative CNN. In order to tackle these problems, we introduce a deep learning network namely expectation-maximization attention cross residual network (EACRN) to tackle the super-resolution task. Particularly, we propose a cross residual in cross residual (CRICR) structure that makes up very deep networks consisting of multiple cross residual groups (CRG) with global residual skip connections. Every cross residual group (CRG) consists of some cross residual blocks with cross short skip connections. At the same time, CRICR allows network focused on capturing high-frequency patterns by connecting rich low-frequency patterns to be bypassed and several short skip connections. In addition, we introduce various convolution kernel size so that adaptive capture the image pattern in different scales, which make these features get the more efficacious image information through interacting with each other. The introduced Expectation-Maximization Attention (EMA) module is robust to the variance of input and is also friendly in memory and computation. Extensive experiments demonstrate our EACRN obtains superior performance and visual effect relative to the most advanced algorithm.

1. Introduction

super-resolution is a basic task in computer vision, especially single-image super-resolution (SISR), which has attracted a lot of interest from researchers. The target of SISR is that produce clear and high-resolution (HR) images, given blurry and low-resolution (LR) images. However, SISR is an ill-posed problem because converting low-resolution images to super-resolution images exists multiple solutions. To tackle this inverse problem, numerous learning method based on big data is widely applied to the learning the mapping from LR images to HR images.

Recently, deep neural networks (DNN) [25, 24, 23, 8] have shown this way can improve significant performance for single image resolution problems. Dong et al. proposed SRCNN [6] in 2014, which is first use CNN to SISR issues. SRCNN is an efficient network that can learn an end to end mapping from LR to HR images and achieve satisfactory performance. Since then, many studies have proposed a large number of CNN models [14, 7, 26, 15, 17, 27, 21, 39] focused on learning the mapping from low-quality images and high-quality images and finding the locally optimal solution. EDSR [21] is the champion in the NTIRE2017 Challenge. It is based on SRResNet [18] and enhances the representational capacity of neural networks by not using batch normalization layer and construct wider and deeper network structures. The great improvements on EDSR have demonstrated the depth of neural network is significant for single image super-resolution problems. These models have reached excellent performance in the SISR problem for SSIM [33] and PSNR. However, to the best of our knowledge, these models tend to simply stack residual blocks to build more complicated connections and network architecture, which means that the training model needs more time, tricks, and resources. Whether the deeper network makes further contributions to SISR and how to build a deeper network remains to be explored.

In order to seek some problems in traditional models, we reproduce some typical models such as SRCNN [6], EDSR [21], SRResNet [18], MSRN [19], RCAN [39], and SAN [5]. In the reproduce experiments, we found that most traditional algorithms exist some problems. Firstly, the change of the subtle network architecture are sensitive to most mod-

*Corresponding authors.

els and some models are hard to obtain the performance of the paper since without network training techniques and configuration, such as data normalization, gradient truncation, and weight initialization, which means the improvement of performance not always due to the changes in the model, but rather some unknown training techniques. Secondly, most methods improve the performance of the network by blindly increase the depth of the network but ignore making full use of the high-frequency information and original low-resolution image of features. With the network deeper, the computation of the network is larger and gradient information gradually disappears since the transmission process. It is critical to make the best use of this high-frequency information for reconstructing super-resolution images. Most existing SR models simply stacking blocks and add residual skip connections to improve performance for reconstructing more clear images. super-resolution task is an ill-pose problem. It should be noted that simple linear mapping does not work very well. In order to capture long-range dependencies, several works develop the non-local block [32]. However, the operation of the non-local block needs to generate a large attention map, which will lead to high computation complexity and occupies a huge number of GPU memory. Although most SR models stacking more and more blocks to enhance the nonlinear mapping for higher performance, this not only made the network cumbersome but also may not achieve the desired effect. How to build a complex network and make full use of the learning ability of the network is important for the model to reconstruct super-resolution images.

To practically resolve the above problems, we rethink the attention mechanism from the view of the expectation-maximization (EM) algorithm [20] and introduce a novel deep learning namely expectation-maximization attention cross residual network (EACRN) to tackle single image super-resolution problem. Besides, we proposed a complicated base module namely multi-scale cross residual block to construct the building architecture for EACRN. In order to stable the training of the deep network, we proposed a standard deep learning architecture namely cross residual in cross residual (CRICR) structure, the cross residual group (CRG), and global residual skip connection (RSC) build basic architecture to guide residual learning and gradient information flow. For each CRG module, stacking multiple cross residual blocks (CRB) with local cross residual skip connection (LCRSC) is a nuclear structure. Global and local cross-residual skip connections, as well as shortcuts in the cross-residual block (CRB), are used to obtain rich low-frequency information through these identification-based skip connections, thus simplifying the flow of information.

2. Related Work

We roughly divided the SISR task into three main stages. Like linear and bicubic interpolation, these early methods based on sampling theory run very fast, but do not recovery textures and details. The goal of the improved work is to create complex mappings for converting low-quality images to high-quality images. These algorithms rely on embedded techniques from neighbors to sparse coding.

2.1. Traditional Method

Single image super-resolution is an unsteadiness inverse problem since super-resolution images can have multiple solutions for high-resolution images. Some traditional super-resolution methods attempt to constrain the solution space with prior information such as neighbor embedding [4], [3], anchored neighborhood regression [30] and sparse coding [36], [35], [34]. In the assumption of [18], low-dimensional non-linear manifolds have a similar local geometry between low-resolution images and high-resolution image pattern space. When the sample is large enough, the weight calculated by the LR feature domain is used to reconstruct the patch of the HR feature domain to the weighted average of local neighbors. With the assumption of the low-resolution image patches shares the same sparse representation with corresponding high-resolution image patches counterparts, Yang et al. [36], [35] proposed an efficient method to resolve the super-resolution task. The improved work such as [10], [9] uses the prior self-similarity that clear image patches in natural images. Huang et al. [12] proposed the SelfExSR. This method makes use of geometric changes to expand the internal image patches search space. Although SelfExSR does not require a training process, is time-consuming because involves the internal image patches search process.

2.2. Deep Neural Network for Image Super-Resolution

Dong et al. [6] did the pioneering work that proposed SRCNN for SISR and obtained outstanding performance against traditional methods. SRCNN has been further enhanced in VDSR [14] and DRCN [15]. These algorithms first interpolate low-resolution image input into the desired size, which greatly augments the computational effort and inevitably loses some detail. Extracting features from low-quality images and increasing resolution at the end of the network is a better method for the deep neural network. In order to speed up the training and testing speed of SRCNN, a faster network structure FSRCNN [7] is proposed. Ledig et al. [18] proposed ResNet with a generative adversarial network (GAN) and perceptual loss [13] for realistic SR [11]. Nevertheless, most of these algorithms have limited network learning ability, which has limited the representation ability of the network.

2.3. Feature Extraction Block for Image Super-Resolution

Nowadays, many researchers concentrate on feature extraction blocks. Kim et al. proposed a residual skip connection structure in 2016 to speed up the training of networks so that they obtain better results. After that, Huang et al. proposed a dense block. The residual blocks and dense blocks use a fix-sized convolution kernel, and the dense blocks have dense skip connections so which increases the computing complexity. In 2018 Li et al. [19] proposed MSRN to capture different scale image features, which introduces different convolution kernels sizes to adaptively extract features in different scales. Although MSRN uses multi-scale convolutional kernels to capture representations, it fails to make full use of the learning ability of the network. To address these shortcomings, we come up with a cross residual block.

2.4. Attention Model for Image Super-Resolution

Attention is important for various tasks such as machine translation, visual question answering, and video classification. The self-attention methods [31] calculate the context coding at one position by a weighted summation of embeddings at all positions in sentences. Zhang et al. introduced the residual channel attention in RCAN [39], but they did not employ the self-attention method. Dai et al. introduced the non-local block [32] in SAN [5] to capture the long-range dependencies, but it's expensive on the GPU.

Our network is motivated by the success of attention in the above works. We rethink the attention mechanism from the view of the EM algorithm and compute the attention map in an iterative manner as the EM algorithm. At the same time, a cross residual skip connection is applied between different batches. In addition, we introduced local residual skip connection to ease training difficulty, which combined with multi-scale extractor so the features pattern can be reused and shared with each other. A more detailed description will be given in section 3.1.

3. Proposed Method

In this section, we describe in detail each main component of our come up with EACRN for SISR. As seen in Fig.1, EACRN is made up of four parts: a shallow feature extraction module, several cross residual groups (CRG), an upsampling layer, and a reconstruction layer.

3.1. Shallow Feature Extraction

In SISR problem, the HR image I_{HR} degenerates an LR image I_{LR} by downsampling and blurring. the degraded LR image I_{LR} can be represented as

$$I_{LR} = DB(I_{LR}) + n \tag{1}$$

where *D* and *B* represent the downsampling and blurring operations, respectively, and *n* denote the additive noise in the degradation operations. Let us denote the I_{LR} and I_{HR} as the observed I_{LR} input and the estimated I_{HR} output of our EACRN.

We use a 3×3 convolution layer to capture the shallow pattern from the original low-quality image in the feature extraction module. The shallow feature x_1 through the first convolutional layer can be represented as

$$F_0 = H_1(I_{LR}) \tag{2}$$

where $H_1(\cdot)$ denotes the first feature extraction operation. x_1 is served as the input of the following state and used for further feature extraction.

3.2. Cross Residual in Cross Residual Structure

Now we introduce our come up with CRICR architecture in Fig. 1. The CRICR architecture includes *G* cross residual groups (CRG) with long skip connection (LSC). Every CRG consists of *B* cross residual blocks (RCAB) with cross short skip connection (CSSC). Our cross residual in cross residual structure can train the deeper model to capture the high-frequency pattern and learn the complex mapping between blurry low-quality images and clear high-quality images for the single image super-resolution task with high performance.

Stacking several blocks and LSC has been demonstrated to construct the deeper network in [21] In computer vision recognition, Stacking multiple residual blocks to construct a very deep network suffer from hardly improve performance again and training difficulty. Inspired by the previous method EDSR [21], we introduce cross residual group (CRG) as a foundational structure for models. We can formulate a CRG in the *g*-th group as

$$F_g = H_g(F_{g-1}) = H_g(H_{g-1}(\cdots H_1(F_0)\cdots))) \tag{3}$$

where H_g represents the operation of *g*-th CRG. F_{g-1} and F_g denote the input and output for *g*-th CRG. We find simply stacking multiple CRGs would not always to obtain better performance. In order to fix this problem, we introduce the long skip connection in CRICR for stabilizing the flow of gradient information and ease the training. LSC can improve performance with cross residual learning via

$$F_{DF} = F_0 + W_{LSC}F_G = F_0 + W_{LSC}H_g(H_{g-1}(\cdots H_1(F_0)\cdots))) \tag{4}$$

where W_{LSC} is the weight matrix set to the convolution layer at the tail of CRICR. For simplicity, we omit the bias term. LSC can not only stabilize the training and slow down the flow of gradient information across CRGs but only make CRICR capture more residual information.

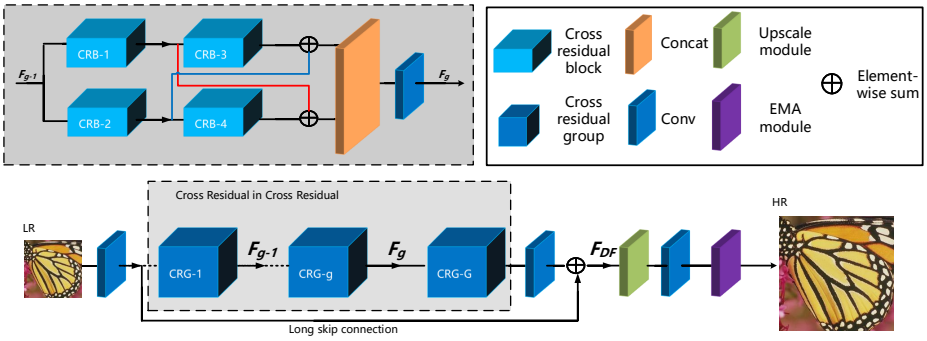

Figure 1. Network architecture of our expectation-maximization attention cross residual network (EACRN)

Since the LR inputs and functions contain a wealth of information, and the SR network aims to reconstruct more useful information. Identity-based skip connections can bypass a large number of low-frequency information. In order to further towards residual learning, we stacked four cross residual blocks in each CRG. The four cross residual blocks (CRBs) in the g-th RG may be represented as

$$F_{g,1} = H_{g,1}(F_{g-1}) \tag{5}$$

$$F_{g,2} = H_{g,2}(F_{g-1}) \tag{6}$$

$$F_{g,3} = H_{g,3}(F_{g,1}) \tag{7}$$

$$F_{g,4} = H_{g,4}(F_{g,2}) \tag{8}$$

$$F_g = W_{conv}([F_{g,1} + F_{g,4}, F_{g,2} + F_{g,3}]) \tag{9}$$

where F_{g-1} and $F_{g,b}$ are the input of the g-*1*-th CRG and output of the b-th CRB in g-th CRG.$[\cdot]$ means the *concat* feature fusion operation, and the W_{conv} denote the weight of convolution layer.

3.3. Cross Residual Block

In order to enhance non-linear mapping for super-resolution reconstruction, we propose cross residual block (CRB). We will elaborate on this structure. As shown in Fig.2, in the head of our CRB we set two branches operated by two 5 convolution layers to largely extract feature, and then we use a multi-scale convolution layer to learn adaptively extracting features. Therefore, the gradient flow information between two bypasses allows detecting image patterns at different scales. We also adopt cross residual learning for our CRB. Since single image super-resolution is an ill-posed problem existing multiple solutions, cross residual skip connection not only enhance the non-linear mapping but also constrains linear condition.

3.4. Expectation-Maximization Attention Module

In order to capture long-range dependencies, we propose an expectation-maximization attention module (EMA). We will elaborate on this structure. As shown in Fig.3, Our proposed EMA consists of three operations, including responsibility estimation (A_E), likelihood maximization (A_M), and data re-estimation (A_R). Briefly, given the input $X \in R^{N\times C}$ and the initial bases $u \in R^{K\times C}$, A_E estimates the latent variables Z $\in R^{N\times K}$, so it functions as the E step in the EM algorithm. A_M uses the estimation to update the bases u, which works as the M step. Then, with the converged u and Z, A_R reconstructs the original X as Y and outputs it.

3.5. Loss Function

According to [21], it may not be possible to use L2 loss restoring sharp edges since L2 loss will result in excessive smoothing. For image SR, the L1 loss function provides better convergence than the L2 loss function. We found that using the L1 training network obtains better performance in PSNR and visually. The quality is lost compared to L1. So we use L1 loss instead of L2. Given the training data sets $\{I_{LR}, I_{HR}\}^N$, where the N denotes the number of the datasets, ground truth HR patch I_{HR}, and the low-resolution LR patch I_{LR}, the loss function with the parameter set Θ is

$$L(\Theta) = \frac{1}{N}\sum_{i=1}^{N} ||I_{HR} - I_{SR}||_1 \tag{10}$$

4. Experiments

4.1. Datasets

In this work, we use DIV2K datasets as training datasets and choose 800 training images from them. [29]. Data augmentation performs on training images for our models, which are randomly rotated and flipped horizontally. SET5 [2], SET14 [37], BSDS100 [1], and URBAN100 [12] were chosen as benchmark datasets for testing and we compare our models with several advanced methods: . The PSNR and the SSIM as performance indicators for all models.

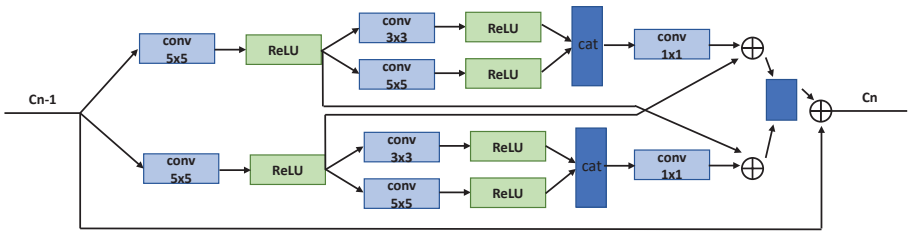

Figure 2. The structure of cross residual block (CRB).

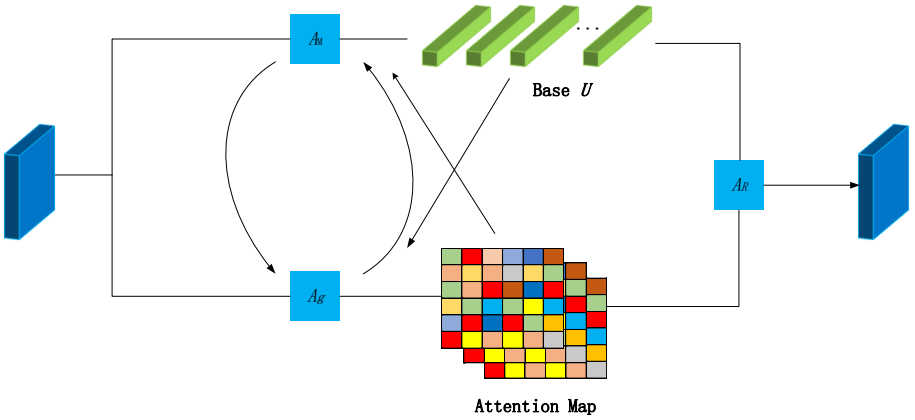

Figure 3. The structure of expectation-maximization module (EMA).

4.2. Implementation Details

In our EACRN, we set the CRG number as 10. In the CRICR structure, each CRB and convolutional layer have 64 filters with the stride of 1, in addition to the upsampling layer and the reconstruction layer. The upsampling layer in our advanced module is sub-pixel convolution [26]. The reconstruction layer is a three-channel of stride 1 of convolution layer to reconstruct the HR image. As for some convolution layers, we use the ReLU function [22] as the activation function.

For the training phase, the original low-resolution images are RGB color space and all channel is processed. Low-resolution image patches are obtained through downsampling the high-resolution image patches adopting Matlab function bicubic interpolation. We randomly sample 16 high-resolution image patches with the size of 192 $\times$ 192 without overlapping in each training batch. In order to demonstrate our method can easily insert any deep learning model as a basic module and improve performance, we did not use the complex weight initial method. We implement EACRN to optimize the models using Adam [16] method as an optimizer with the PyTorch package. We set the momentum parameter as 0.9 and initialize the learning rate as $1e$-4 and decreased by half for every 200 epochs. The training of a single EACRN model can roughly take 7 days with a GTX 1080Ti GPU.

Table 1. Effects of different modules with cross residual learning. We report the best PSNR and SSIM values on Set5 images.

Method	Cross residual learning				
CRG		✓			✓
CRB			✓		✓
EMA				✓	✓
PSNR	31.23	31.45	31.87	32.16	32.23

4.3. Ablation Study

In this subsection, we analyze the effects of the cross residual learning in our CRG and CRB. We now perform analyses on the proposed cross residual learning for CRG and CRB in detail, i.e., the cross residual group (CRG) with cross residual learning and cross residual block with cross residual learning and the no cross-learning from the above two components. We design five networks that set the same numbers of feed-forward features in CRB and the same numbers of CRG to build the standard model. The model

Table 2. Quantitative comparisons of state-of-the-art methods.

Method	Scale	Set5	Set14	BSD100	Urban100	Manga109
Bicubic	2	33.66/.9299	30.24/.8688	SSIM	26.88/.8403	30.80/.9339
SRCNN	2	36.66/.9542	32.45/.9067	31.36/.8879	29.50/.8946	35.60/.9663
VDSR	2	37.53/.9590	33.05/.9130	31.90/.8960	30.77/.9140	37.22/.9750
LapSRN	2	37.52/.9591	33.08/.9130	31.08/.8950	30.41/.9101	37.27/.9740
MemNet	2	37.78/.9597	33.28/.9142	32.08/.8978	31.31/.9195	37.72/.9740
EDSR	2	38.11/.9602	33.92/.9195	32.32/.9013	32.93/.9351	39.10/.9773
SRMD	2	37.79/.9601	33.32/.9159	32.05/.8985	31.33/.9204	38.07/.9761
DBPN	2	38.09/.9600	33.85/.9190	32.27/.9000	32.55/.9324	38.89/.9775
RDN	2	38.24/.9614	34.01/.9212	32.34/.9017	32.89/.9353	39.18/.9780
RCAN	2	38.27/.9614	34.11/.9216	32.41/.9026	33.34/.9384	39.43/.9786
SAN	2	38.31/.9620	34.07/.9213	32.42/.9028	33.10/.9370	39.32/.9792
EACRN	2	38.42/.9632	34.17/.9233	32.47/.9038	33.12/.9350	39.42/.9893
Bicubic	3	39.32/.9792	27.55/.7742	27.21/.7385	24.46/.7349	26.95/.8556
SRCNN	3	32.75/.9090	29.30/.8215	28.41/.7863	26.24/.7989	30.48/.9117
VDSR	3	33.67/.9210	29.78/.8320	28.83/.7990	27.14/.8290	32.01/.9340
LapSRN	3	33.82/.9227	29.87/.8320	28.82/.7980	27.07/.8280	32.21/.9350
MemNet	3	34.09/.9248	30.01/.8350	28.96/.8001	27.56/.8376	32.51/.9369
EDSR	3	34.65/.9280	3.52/ .8462	29.25/.8093	28.80/.8653	34.17/.9476
SRMD	3	34.12/.9254	30.04/.8382	28.97/.8025	27.57/.8398	33.00/.9403
RDN	3	34.71/.9296	30.57/.8468	29.26/.8093	28.80/.8653	34.13/.9484
RCAN	3	34.74/.9299	30.64/.8481	29.32/.8111	29.08/.8702	34.43/.9498
SAN	3	34.75/.9300	30.59/.8476	29.33/.8112	28.93/.8671	34.30/.9494
EACRN	3	34.85/.9321	30.63/.8576	29.54/.8121	28.99/.8771	34.41/.9497
Bicubic	4	28.42/.8104	26.00/.7027	25.96/.6675	23.14/.6577	24.89/.7866
SRCNN	4	30.48/.8628	27.50/.7513	26.90/.7101	24.52/.7221	27.58/.8555
VDSR	4	31.35/.8830	28.02/.7680	27.29/.0726	25.18/.7540	28.83/.8870
LapSRN	4	31.54/.8850	28.19/.7720	27.32/.7270	25.21/.7560	29.09/.8900
MemNet	4	31.74/.8893	28.26/.7723	27.40/.7281	25.50/.7630	29.42/.8942
EDSR	4	32.46/.8968	28.80/.7876	27.71/.7420	26.64/.8033	31.02/.9148
SRMD	4	31.96/.8925	28.35/.7787	27.49/.7337	25.68/.7731	30.09/.9024
DBPN	4	32.47/.8980	28.82/.7860	27.72/.7400	26.38/.7946	30.91/.9137
RDN	4	32.47/.8990	28.81/.7871	27.72/.7419	26.61/.8028	31.00/.9151
RCAN	4	32.62/.9001	28.86/.7888	27.76/.7435	26.82/.8087	31.21/.9172
SAN	4	32.64/.9003	28.92/.7888	27.78/.7436	26.79/.8068	31.18/.9169
EACRN	4	32.67/.90021	28.95/.7894	27.81/.7456	26.87/.8087	31.23/.9188

$MBASE$ is obtained by removing CRG , CRB, and EMA that is based on the basic EACRN, which be made up of the standard framework. The performance (PSNR = 31.23 dB) of M_BASE is bad that is caused by the hard of train-

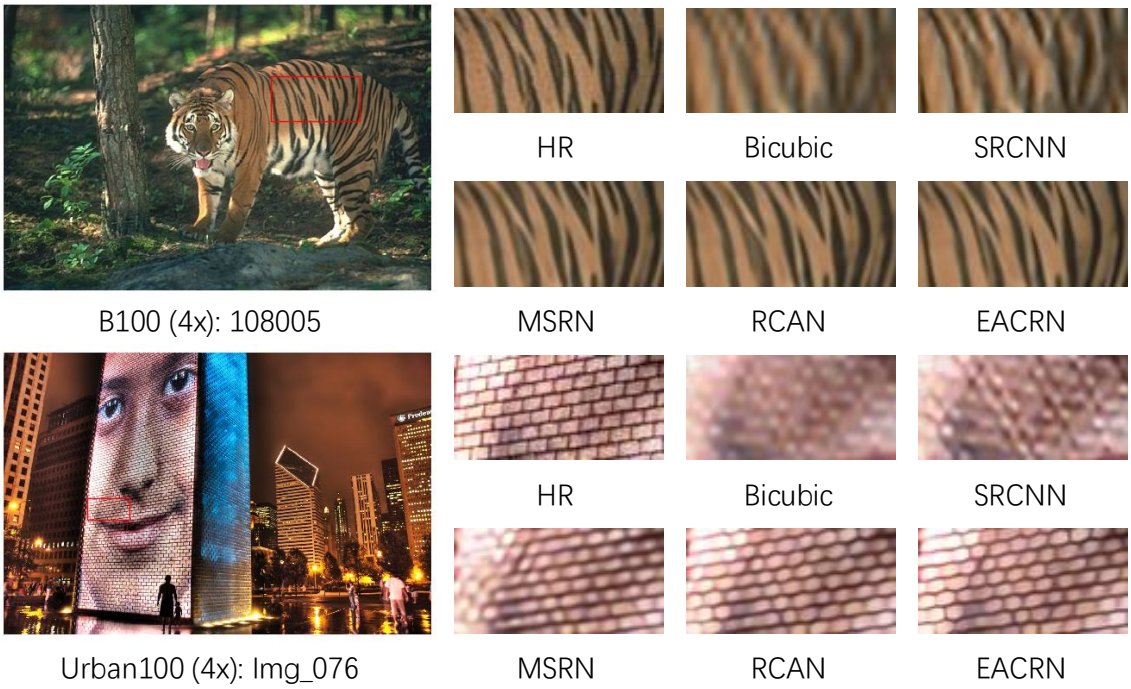

Figure 4. Visual comparison for 4× SR with BI model on benchmark dataset.

Table 3. Specifications comparison with RCAN.

Algorithm	Feature extraction	Filters	Parameters	Updates
RCAN	10 blocks	64	16M	4×10^5
EACRN	10 blocks	64	15.4M	1×10^3

ing and inefficient flow of gradient information. It demonstrates that stacking several basic convolution layers does not obtain better performance. Then, we add cross residual learning to the CRG and CRB of M_BASE to produce M_CRG and M_CRB. The results demonstrate each structure can improve the performance of M_BASE. The main reason is largely due to that cross skip connection allows more abundant low-frequency information from the LR images to be bypassed. The combination of CRB, CRG, and EMA will improve performance than it. When we both use cross residual learning in CRG, CRB, and EMA, the performance can be further improved.

4.4. Comparisons with State-of-the-art Methods

We choose many state-of-the-art SR models to compare with our model, including Bicubic, SRCNN [6], VDSR [14], LapSRN [17], MemNet[28], EDSR [21], SRMD [38], DBPN[13], RDN[40], RCAN[39], and SAN[5]. All the object quantitative results for different scale factors are reported in Table 2. Compared with other algorithms, our EACRN performs better results on all the datasets on various scaling factors. This is mainly because both of them cross residual learning to learn feature interdependencies so which prompts the network to focus on more informative features. Compared with MSRN, our EACRN obtains better results for datasets (e.g., such as Set5, Set14, and BSD100) with rich texture information, while obtaining a little worse results for datasets(e.g., Urban100 and Manga109) with rich repeated edge information. It is well known that textures have more complex statistical properties and are high-order patterns, while edges are first-order patterns that can be extracted by a step operator. Therefore, our EACRN based on cross residual learning to excavate second-order feature statistics works better on images with more high-order information (such as texture).

For the purpose of proving the subjective effect of our proposed model, we also show the scaling results for the different methods for visual comparisons on the Urban100 and

Set14 data sets for 4 × SR in Figure 3. From this, we can see that many compared SR models do not accurately reconstruct the texture and suffer from severe fuzzy artifacts. In contrast, our EACRN gets clearer results and can recover more high-frequency details such as high contrast and sharp edges. Taking "img_005" as an example, most comparison methods output severe fuzzy artifacts. The traditional algorithms bicubic, SRCNN even lost the main structure. Our EACRN can restore the main outline and restore more image detail. Compared with the ground truth, EACRN achieves more faithful results, reconstructs more image details and EACRN has sharper results. These observations verify the superiority of EACRN with more powerful representational ability. Although it is difficult to recovery high-frequency information due to limited information available in low-resolution image input (scaling factor 4 and 8), our EACRN can also make the best use of the limited low-quality information through cross skip connections for more powerful feature expressions so that produce finer results.

5. Conclusions

In this paper, we introduce a new attention mechanism, namely the expectation-maximization attention (EMA), and present a deep expectation-maximization attention expectation-maximization attention cross residual network (EACRN) to tackle the image SR problem. The experiments have demonstrated EMA is robust to the variance of input. In particular, the cross residual in cross residual architecture allows EACRN to capture structural information by embedding cross skip connection operations into the network. At the same time, CRG allows a large number of low-frequency information in the LR image to be bypassed by the cross skip connection. For feature correlation, we propose CRB to learn the interdependence of features to achieve a more discriminative representation. Extensive experiments on SR through the BI model demonstrate our EACRN obtain superior performance against some state-of-the-art models in terms of quantitative and visual results.

References

[1] Pablo Arbelaez, Michael Maire, Charless Fowlkes, and Jitendra Malik. Contour detection and hierarchical image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 33(5):898–916, 2010. 4

[2] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie Line Alberi-Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. 2012. 4

[3] Hong Chang, Dit-Yan Yeung, and Yimin Xiong. Super-resolution through neighbor embedding. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 1, pages I–I. IEEE, 2004. 2

[4] Xiaoxuan Chen and Chun Qi. Low-rank neighbor embedding for single image super-resolution. *Signal Processing*, 94(1):6–22, 2014. 2

[5] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11065–11074, 2019. 1, 3, 7

[6] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2015. 1, 2, 7

[7] Chao Dong, Chen Change Loy, and Xiaoou Tang. Accelerating the super-resolution convolutional neural network. In *European conference on computer vision*, pages 391–407. Springer, 2016. 1, 2

[8] X. Du, S. Jiang, Y. Si, L. Xu, and C. Liu. Mixed high-order non-local attention network for single image super-resolution. *IEEE Access*, 9:49514–49521, 2021. 1

[9] Gilad Freedman and Raanan Fattal. Image and video upscaling from local self-examples. *ACM Transactions on Graphics (TOG)*, 30(2):12, 2011. 2

[10] Daniel Glasner, Shai Bagon, and Michal Irani. Super-resolution from a single image. In *2009 IEEE 12th international conference on computer vision*, pages 349–356. IEEE, 2009. 2

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 2

[12] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5197–5206, 2015. 2, 4

[13] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 2, 7

[14] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1646–1654, 2016. 1, 2, 7

[15] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Deeply-recursive convolutional network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1637–1645, 2016. 1, 2

[16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[17] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 624–632, 2017. 1, 7

[18] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken,

Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017. 1, 2

[19] Juncheng Li, Faming Fang, Kangfu Mei, and Guixu Zhang. Multi-scale residual network for image super-resolution. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 517–532, 2018. 1, 3

[20] Xia Li, Zhisheng Zhong, Jianlong Wu, Yibo Yang, Zhouchen Lin, and Hong Liu. Expectation-maximization attention networks for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9167–9176, 2019. 2

[21] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 136–144, 2017. 1, 3, 4, 7

[22] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010. 5

[23] Ben Niu, Weilei Wen, Wenqi Ren, Xiangde Zhang, Lianping Yang, Shuzhen Wang, Kaihao Zhang, Xiaochun Cao, and Haifeng Shen. Single image super-resolution via a holistic attention network. In *European Conference on Computer Vision*, pages 191–207. Springer, 2020. 1

[24] Wenqi Ren, Sifei Liu, Lin Ma, Qianqian Xu, Xiangyu Xu, Xiaochun Cao, Junping Du, and Ming-Hsuan Yang. Low-light image enhancement via a deep hybrid network. *IEEE Transactions on Image Processing*, 28(9):4364–4375, 2019. 1

[25] Wenqi Ren, Lin Ma, Jiawei Zhang, Jinshan Pan, Xiaochun Cao, Wei Liu, and Ming-Hsuan Yang. Gated fusion network for single image dehazing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3253–3261, 2018. 1

[26] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. 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*, pages 1874–1883, 2016. 1, 5

[27] Ying Tai, Jian Yang, and Xiaoming Liu. Image super-resolution via deep recursive residual network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3147–3155, 2017. 1

[28] Ying Tai, Jian Yang, Xiaoming Liu, and Chunyan Xu. Memnet: A persistent memory network for image restoration. In *Proceedings of the IEEE international conference on computer vision*, pages 4539–4547, 2017. 7

[29] Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, and Lei Zhang. Ntire 2017 challenge on single image super-resolution: Methods and results. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 114–125, 2017. 4

[30] Radu Timofte, Vincent De, and Luc Van Gool. Anchored neighborhood regression for fast example-based super-resolution. In *IEEE International Conference on Computer Vision*, 2014. 2

[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 3

[32] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018. 2, 3

[33] Zhou Wang, Alan C Bovik, Hamid R Sheikh, Eero P Simoncelli, et al. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 1

[34] Jianchao Yang, Zhaowen Wang, Zhe Lin, Scott Cohen, and Thomas Huang. Coupled dictionary training for image super-resolution. *IEEE transactions on image processing*, 21(8):3467–3478, 2012. 2

[35] Jianchao Yang, John Wright, Thomas Huang, and Yi Ma. Image super-resolution as sparse representation of raw image patches. In *2008 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2008. 2

[36] Jianchao Yang, John Wright, Thomas S Huang, and Yi Ma. Image super-resolution via sparse representation. *IEEE transactions on image processing*, 19(11):2861–2873, 2010. 2

[37] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *International conference on curves and surfaces*, pages 711–730. Springer, 2010. 4

[38] Yun Zhang. Problems in the fusion of commercial high-resolution satelitte as well as landsat 7 images and initial solutions. *International Archives of Photogrammetry Remote Sensing and Spatial Information Sciences*, 34(4):587–592, 2002. 7

[39] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 286–301, 2018. 1, 3, 7

[40] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2472–2481, 2018. 7