Neural Computing and Applications
https://doi.org/10.1007/s00521-019-04609-8

DEEP LEARNING APPROACHES FOR REALTIME IMAGE SUPER RESOLUTION (DLRSR)

CASR: a context-aware residual network for single-image super-resolution

Yirui Wu[1,2] · Xiaozhong Ji[2] · Wanting Ji[3] · Yan Tian[4] · Helen Zhou[5]

Received: 2 April 2019 / Accepted: 8 November 2019

Abstract

With the significant power of deep learning architectures, researchers have made much progress on super-resolution in the past few years. However, due to low representational ability of feature maps extracted from nature scene images, directly applying deep learning architectures for super-resolution could result in poor visual effects. Essentially, unique characteristics like low-frequency information should be emphasized for better shape reconstruction, other than treated equally across different patches and channels. To ease this problem, we propose a lightweight context-aware deep residual network named as CASR network, which appropriately encodes channel and spatial attention information to construct context-aware feature map for single-image super-resolution. We firstly design a task-specified inception block with a novel structure of astrous filters and specially chosen kernel size to extract multi-level information from low-resolution images. Then, a Dual-Attention ResNet module is applied to capture context information by dually connecting spatial and channel attention schemes. With high representational ability of context-aware feature map, CASR can accurately and efficiently generate high-resolution images. Experiments on several popular datasets show the proposed method has achieved better visual improvements and superior efficiencies than most of the existing studies.

Keywords Context-aware residual network · Channel and spatial attention scheme · Inception block · Single-image super-resolution

1 Introduction

Super-resolution (SR) is generally defined as a process to obtain high-resolution (HR) images form inputs of low-resolution (LR) observations. There exists a rough but classical taxonomy way to category SR methods based on number of input LR images: single-image super-resolution (SISR) and multiple images super-resolution (MISR). Being a highly ill-posed problem, SISR is more challenging than MISR, since it requires to hallucinate missing image details by learning the relationship between LR and HR from a training dataset.

✉ Wanting Ji
Wanting.ji@massey.ac.nz

Yirui Wu
wuyirui@hhu.edu.cn

Xiaozhong Ji
shawn_ji@163.com

Yan Tian
tianyan@zjgsu.edu.cn

Helen Zhou
helen.zhou@manukau.ac.nz

[1] College of Computer and Information, Hohai University, Nanjing, China

[2] National Key Lab for Novel Software Technology, Nanjing University, Nanjing, China

[3] School of Natural and Computational Sciences, Massey University, Auckland, New Zealand

[4] Zhejiang Gongshang University, Hangzhou, China

[5] School of Engineering, Manukau Institute of Technology, Auckland, New Zealand

Published online: 20 November 2019

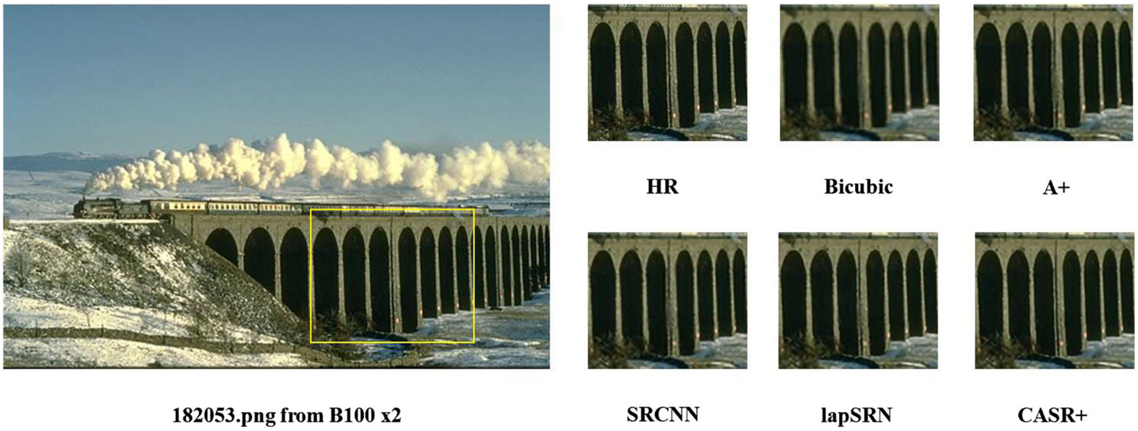

Fig. 1 Comparisons on visual effects among CASR+, Bicubic, A+ [42], SRCNN [8] and LapSRN [23] with 2× scale factor. It's noted CASR+ is an improved version of CASR by utilizing self-ensemble property during testing. We enlarge the patch labeled by yellow rectangle to offer a better visual comparison, where we can clearly observe poor visual effects like blur, geometry preservation and detail reconstruction failures achieved by comparative methods. Meanwhile, we can notice CASR+ maintains intrinsic geometry information like lines or etc, and clearly reconstructs high-frequency details like textures or etc

With the development of deep learning techniques, more works involve convolutional neural network (CNN) [21], generative adversarial network (GAN) [12] or other deep learning models [5] to perform SISR tasks. However, these methods generally suffer from drawbacks in visual effects, which are shown in Fig. 1 by testing with an example sampled from B100 dataset [28]. Based on visual comparisons between groundtruth HR images and HR images produced by various methods, we can observe unpleasant artifacts like blurry patches, failure in reconstructing high-frequency image details and loss of low-frequency features like straight lines or etc. In fact, unpleasant visual effects couldn't be avoided by most of deep learning methods, since their generated feature map is short of context information to capture unique characteristics of input LR images. Without context information, deep-learning-based methods perform the same operation across channels and patches, which loses emphasis on reconstructing task-specified feature map for a better representation.

To deal with this drawback, several attention schemes [15, 58] are recently adopted to recalibrate the extracted feature maps, so that builded networks are more adaptive and sensitive to restore high-frequency features. For example, Zhang et al. [58] adopt an existing channel attention mechanism to construct very deep residual channel attention networks (RCAN). Hu et al. [15] propose a channel-wise and spatial feature modulation (CSFM) network built by feature modulation memory (FMM) modules with stack connections, which successfully transforms low-resolution features to high informative features. However, their adopted attention scheme is either out of modification to be suitable for SISR problem, or lack of emphasis on locally preserving low-frequency features due to their stack structures. Therefore, how to construct an appropriate attention scheme to describe context information for SISR problem remains unresolved. Based on these considerations, we propose to build a context-aware and light-scale deep residual network, named as CASR, to perform SISR task.

The contributions of this paper are threefold:

- We propose a novel deep and context-aware residual network for accurate SISR task, in which a task-specified inception block and DRM structure are adopted to enhance feature representative ability on the basis of multi-level features and context information.
- DRM is carefully designed with a dual combination form between channel and spatial attention scheme. DRM adaptively rescales feature map by exploiting inter-channel and intra-channel interdependencies, thus involving inherent and unique context characteristics of LR images for better reconstruction results. Since DRM only increases a small amount of computation burden and can be easily implemented, we believe it can be adopted by other computer vision tasks with modifications.
- CASR successfully builds a lightweight SISR system with features of high reconstruction accuracy, fast computing speed and only 40M storage size. (For comparison, EDSR [25] reports 120M storage request.) We believe these properties make CASR appropriate to be applied in most of application scenarios, like cellphones, drones or other embedding systems

The rest of the paper is organized as follows. Section 2 gives an overview of the related work on relative aspects. In Sect. 3, details of the proposed CASR network, including network architecture design, Inception block and

DRM structure, is discussed. Section 4 shows our experimental results with several comparative methods, and finally Sect. 5 concludes the paper.

2 Related work

Existing methods related to our work can be categorized into the following two categories: deep learning methods for SISR and attention model.

2.1 Deep learning methods for SISR

In this subsection, we category deep learning methods for SISR into three types, i.e., convolutional neural network (CNN), generative adversarial network (GAN) and other deep learning methods.

CNN-based SISR methods are quite larger in amount than the other two categories of methods, due to their long history and impressive HR reconstruction effects. The first work to solve SISR problem, i.e., SRCNN, is introduced by Dong et al. [8]. Three-layer SRCNN network directly learns an end-to-end mapping between input LR images and the corresponding HR output images. Inspired by success of ResNet, Kim et al. [18] propose very deep convolutional networks (VDSR), which constructs a very deep network with global residual learning and layer stack technologies. VDSR successfully recovers high-frequency details and has impressive property of fast convergence. Following the idea of residual learning, Tai et al. [41] propose deeply recursive residual network (DRRN), which not only uses global residual learning in the identity branch, but also introduces new concept of local residual learning in local residual branch. Involving both global and local residual learning results in desirable visual effects and high performance in measurements.

To pursue better feature map representation, Tong et al. [44] combine low-level and high-level features in a reasonable way by propagating feature maps of each layer into all subsequent layers and allowing dense skip connection. Recently, Lim et al. [25] develop an enhanced deep super-resolution network (EDSR), which generates high-quality feature map by removing unnecessary modules in conventional residual networks and expands model depth with a stable training procedure. Haris et al. [13] propose deep back-projection networks (DBPN), which exploits iterative up- and downsampling layers to provide an error feedback mechanism for less projection errors at each stage. Shamsolmoali et al. [39] present an effective model based on progressive dilated densely connected, and a novel activation function has a nonlinear learnable function with some short connections. These strategies help the network to obtain deep and complex features, which supports the exponential growth of the receptive field, parallel by increasing the filter size.

With the development of edge computing [34, 51], cloud computing [31, 36], big data technology [35, 50], internet of things [32, 48], and other technologies [33, 52], more technologies are adopted to improve efficiency of CNN-based SISR methods. Inspired by the light-scale and effective network design in EDSR [25], we aim to propose a light CNN structure, which is fast in running speed and small in storage size.

Due to unsupervised training feature of GAN, GAN-based SISR methods could handle a large amount of unlabeled images without any prior knowledge on inputting LR and HR images. Super-resolution generative adversarial network (SRGAN) [24] is first proposed to apply GAN model on SISR task, which is trained under the constraint of perceptual similarity. Specifically, perceptual similarity is defined as a sum value of adversarial loss and content loss, where the former one is specially designed to guarantee that SRGAN could generate high-quality and photo-realistic HR images with help of a discriminator network. Afterward, Bulat et al. [3] propose a two-stage process, which uses a GAN to learn how to perform image degradation at first and then learn image super-resolution with the trained GAN. To pursue high quality for large upsampling factors, Wang et al. [45] propose ProGANSR, which is progressive in both architecture and training: the network upsamples an image in intermediate steps, while the learning process is organized from easy to hard. Most recently, Shamsolmoali et al. [40] organize a GAN-based model for SR tasks by a gradual learning process from simple to advanced, which means from the small upsampling factors to the large upsampling factor that improves the overall stability of the training.

Deep Reinforcement Learning (DRL) recently has been introduced for SISR task. Following the idea of reinforce learning, DRL for SISR utilizes reward scheme to navigate up-scaling regions, which results in an adaptively optimizing way. For instance, Cao et al. [5] propose a novel attention-aware Face Hallucination framework, which first follows principles of DRL to sequentially discover patches required to up-scale and then exploits global characteristics of inputting facial image to enhance facial patch. Moreover, Cao et al. [4] propose a novel SR method with multi-channel constraints learning conception, which integrates clustering, collaborative representation, and progressive multilayer mapping relationships to reconstruct high-resolution (HR) color image.

2.2 Attention model

Attention model sources from visual attention mechanism found in humans. Human generally intends to focus on part of scene over time to obtain important and informative messages to comprehend his or her surroundings. Based on study from human brain, attention model, regarded as an automatically and selectively focusing mechanism, has been deployed in various deep-learning-based applications. The key effects to apply attention model lies in two aspects, namely, decide meaningful parts of input to focus on and allocate limited computing resources to important parts for higher efficiency. Based on different mechanisms, researchers generally classify current attention models into two categories, i.e., hard and soft attention, where the former one selects certain parts of input signal to focus on, and the latter one assigns different weights to parts of input signal for selection.

Hard attention is firstly introduced by Mnit et al. [29], which adaptively selects a sequence of regions as informative parts according to a group of criterions. These selected regions are then regarded as input for later RNN network to perform recognition or other tasks. Following the idea to focus on informative parts of input data, He et al. [14] propose a novel text-attentional convolutional neural network (Text-CNN) for scene text detection, which particularly focuses on text-related features extracted from salient regions based on context information of image components.

Due to its flexibility and efficiency, soft attention is more widely used than hard attention in deep learning structures. Soft attention model is generally formed as a dimension of interpretability into internal representations by selectively focusing on specific information. Specifically, we could divide core procedures of soft attention model into two stages, i.e., calculate weights based on similarity between input signal and pre-trained weights, and re-weight original values based on calculated weights.

During the first stage, multilayer perception (MLP) is often utilized to calculate similarity or correlation between input signal Q and one of the pre-trained weights W_i as $\mathrm{sim}(Q, W_i) = \mathrm{MLP}(Q, W_i)$. Afterward, researchers often adopt Softmax function to perform normalization on calculated similarity and emphasize informative parts based on its inherent ability:

$$\alpha_i = \mathrm{softmax}(\mathrm{sim}(Q, W_i)) = \frac{e^{\mathrm{sim}(Q,W_i)}}{\sum_{j=1}^{L} e^{\mathrm{sim}(Q,W_i)}}; \tag{1}$$

where L refers to the number of pre-trained weights.

In the second stage, attention value *Atten* can be obtained by summing weighted original values with:

$$\mathrm{Atten} = \sum_{i=1}^{L} \alpha_i \cdot v_i \tag{2}$$

where v_i refers to original values and operation · means element-wise operation. By calculating with two stages above, we can get the attention value *Atten* for original vector v with the input signal Q.

Based on the core procedures of soft attention model, various methods are proposed to solve different problems. With the idea of utilizing attention-based weight scheme to fuse information, Yeung et al. [53] propose a sliding window to capture a range of frames as input, which are further assigned with frame-wise attention weights learned by an auto-encoder network. Chen et al. [6] fuse attention with convolutional neural network (CNN) and RNN to automatically extract the most salient modality-specific features, which are further converted to higher level representation for the purpose of human activity recognition with imbalanced labeled data over classes. Anderson et al. [1] propose a combined bottom-up and top-down attention mechanism based on Faster R-CNN, which enables attention to be calculated at the level of objects and other salient image regions. They further apply their novel attention model on image captioning, which results in a new state-of-the-art performance on public datasets. Latest, Zhao et al. [60] propose end-to-end Recurrent Attention (RA) models for pedestrian attribute recognition, which combines the Recurrent Learning and Attention Model to highlight the spatial position on feature map and mine the attention correlations among different attribute groups to obtain more precise attention.

Most related to SR work, Kim et al. [20] propose a novel channel and spatial attention mechanism specially optimized for SR, which prefer to fuse spatial and channel attention for a unity representation before assigning weights, rather than two separate weight schemes. However, their work is only tested on two simplified attention schemes. Woo et al. [47] construct Convolutional Block Attention Module (CBAM) as a lightweight and general attention module, which sequentially infers attention maps along spatial and channel dimensions at first and then multiplies attention maps to the input feature map for adaptive feature refinement. Their proposed light-scale attention module has achieved excellent performance in lots of recognition and classification tasks. Inspired by attention schemes applied in other domains and related SISR work based on attention scheme, we combine channel and spatial attentions in a dual form to construct lightweight DRM structure, which adaptively modulates feature representations for accurate SR with context information among feature channels and different regions.

3 The proposed method

We firstly design a light-scale network architecture to complete general tasks of SISR, i.e., generating feature maps and up-scaling. Then, we design a task-specified inception block to enhance representative ability of generated feature map with multi-branch convolutional layers. Afterward, we propose a novel Dual-Attention ResNet module (DRM) to construct context descriptors for feature map enhancement. Finally, we describe structure details of the proposed channel and spatial attention schemes, which are adopted to construct DRM.

3.1 Network architecture design

The fundamental goal of SISR methods is to hallucinate missing detailed information of super-resolved images. In the literature, SISR is an inherently ill-posed problem, since the informative information contained in LR images is often insufficient to complete the task of reconstruction. There usually exist two tasks for traditional SISR methods: upsampling of LR images to increase image resolution, and removing artifacts including blur and noise. Owing to the significant multi-task ability of deep neural networks, multiple tasks or intentions of SISR can be accomplished by a single neural network:

$$I_H = F(I_L) \tag{3}$$

where I_L and I_H represent input LR and output HR image, respectively, and function $F(\cdot)$ refers to single neural network to accomplish SISR task.

As shown in Fig. 2, the proposed CASR network mainly consists of five parts: input layer, inception block, Dual-Attenion module, upsample layer and output layer. It's noted that all kernels of filters adopted by five blocks represented in Fig. 2 are defined with $3 * 3$ in kernel size. In the first input layer block, we use one convolutional layer (Conv) to extract initial and shallow feature F_{S} for further processing:

$$F_{\mathrm{S}} = H_{\mathrm{S}}(I_L) \tag{4}$$

where function $H_{\mathrm{S}}(\cdot)$ denotes convolution operation of the input layer. F_{S} is then enhanced by inception block with multi-branch convolutional layers:

$$F_{\mathrm{I}} = H_{\mathrm{I}}(F_{\mathrm{S}}) \tag{5}$$

where function $H_{\mathrm{I}}(\cdot)$ indicates multi-branch operations of the proposed inception block.

Afterward, we construct n Dual-Attention ResNet modules (DRMs) to perform task of enhancing feature map with context information, where n is settled as 16 for all tests in the paper. Specifically, the first DRM is employed to generate context feature F_{D1} based on F_{I}:

$$F_{D_1} = H_{D_1}(F_{\mathrm{I}}) = C_1(F_{\mathrm{I}}) + S_1(F_{\mathrm{I}}) + F_{\mathrm{I}} \tag{6}$$

where function $C_1(\cdot)$ and $S_1(\cdot)$ represent operations of channel and spatial attention scheme in the first DRM, respectively, operator $+$ refers to a dual combination form, and the last term of F_{I} represents a stack connection. It's noted that we stack DRM with short skip connection (SSC) for fast convergency. In fact, the long and short skip connection as well as the shortcut in residual block allows abundant low-frequency information to be bypassed through these identity-based skip connections, which can ease the flow of information.

Regarding DRM as the basic module of CASR network, we could construct deep enough network by cascading n DRMs, which can be represented as

$$F_{D_n} = H_{D_n}(F_{D_{n-1}}) = C_n(F_{D_{n-1}}) + S_n(F_{D_{n-1}}) + F_{D_{n-1}} \tag{7}$$

where the last term of $F_{D_{n-1}}$ represents a short stack connection. After processing of n DRM, CASR network performs a Conv, an upsampling operation and an output layer to obtain output HR image:

$$I_H = H_{\mathrm{O}}(H_{\mathrm{U}}(H_{\mathrm{C}}(F_{D_n})) + F_{\mathrm{I}}) \tag{8}$$

where function $H_{\mathrm{C}}(\cdot)$, $H_{\mathrm{U}}(\cdot)$ and $H_{\mathrm{O}}(\cdot)$ represent Conv, upsampling and operations of output layer, respectively, and the last term of F_{I} represents a long skip connection. In upsampling layer, convolutional filter is followed by a pixel-shuffle operation, which enlarges size of feature extracted from Conv filter.

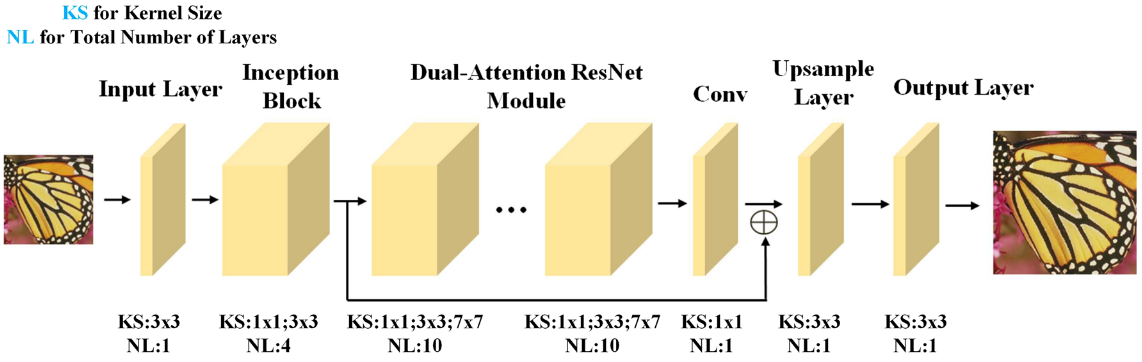

Fig. 2 Network architecture of the proposed CASR network, which consists of input layer, inception block, Dual-Attention module, upsample layer and output layer

Following the definition of SISR task in Eq. 3, CASR can thus be represented as

$$I_H = F_{\mathrm{CASR}}(I_L) = H_{\mathrm{O}}(H_{\mathrm{U}}(H_{\mathrm{C}}(H_{D_n}(\cdots H_{D_1}(H_{\mathrm{I}}(H_{\mathrm{S}}(I_L)))\cdots) + H_{\mathrm{I}}(H_{\mathrm{S}}(I_L)))) \tag{9}$$

During training, the proposed CASR network is designed to optimize with a loss function. To achieve desirable reconstruction results of HR images, we have investigated several loss functions, such as L_1 and L_2 form, perceptual and adversarial losses. We choose L_1 form of loss function based on two considerations. Firstly, it's fair for comparisons by adopting L_1 form since most of residual-based SISR methods use L_1 form for optimization. Secondly, usage of L_1 form improves performance on SSIM after our experiments with different loss functions.

Supposing a training set with N pairs of LR images and corresponding HR images represented as $\{I_L^i, I_H^i; i = 1..,N\}$, the L_1 form of loss function to train the proposed CASR network could be represented as

$$L(\theta) = \frac{1}{N}\sum_{i=1}^{N} \parallel F_{\mathrm{CASR}}(I_L^i, \theta) - I_H^i \parallel \tag{10}$$

where θ denotes the parameter set of our network. We utilize algorithm of stochastic gradient descent to minimize Eq. 10. More details of training parameters and settings can be viewed in Sect. 4.4.

In order to accelerate training and improve the final performance on SISR, we utilize a pre-training strategy for training process. For example, to train CASR for upsampling factor ×3 and ×4, we initialize the model parameters with pre-trained network, of which upsampling factor is settled as ×2. Essentially, such initialization strategy makes CASR converge much faster than start training with random initialization.

3.2 Task-specified inception block

Inspired by [27] to enhance feature representation for better object detection results, we believe it's essential to offer abundant and enhanced low-frequency features for high-frequency detail reconstruction. We thus design a task-specified inception block to perform low-frequency feature map enhancement.

Structure of the proposed inception block is shown in Fig. 3a, where we can notice it contains multi-branch convolutional layers to help capture abundant and variant information from inputting LR image. There exist two main differences between standard and the proposed inception block. Firstly, we utilize special chosen kernel sizes, i.e., 1×1, 3×3, to be cores of multi-branch convolutional layers, which copes with multi-scale property of low-frequency features, thus successfully capturing characteristics of different receptive fields. Secondly, we use two kinds of convolutional layers, i.e., normal and astrous convolutional layers. In fact, astrous convolutional layer is designed to capture information with a larger area, while keeping the number of parameters unchanged. We thus utilize astrous convolutional layer to enlarge receptive field and keep low computation cost at the same time. It's noted the proposed inception block concatenates features generated by multi-branch convolutional layers, which fuses information captured by different receptive fields for a unity and enhanced representation.

The proposed inception block could simulate the relationship between size and eccentricity of receptive fields in human visual systems. In other words, the proposed inception block ensures that positions near the center have larger weights than faraway ones with variety of kernels. In that way, we enhance representative ability of low-frequency features by considering a general rule of human visual systems, i.e., relationship between size and eccentricity. Above all, the proposed inception block successfully enhances low-frequency features with information-abundant, multi-level, and visually-featured properties,

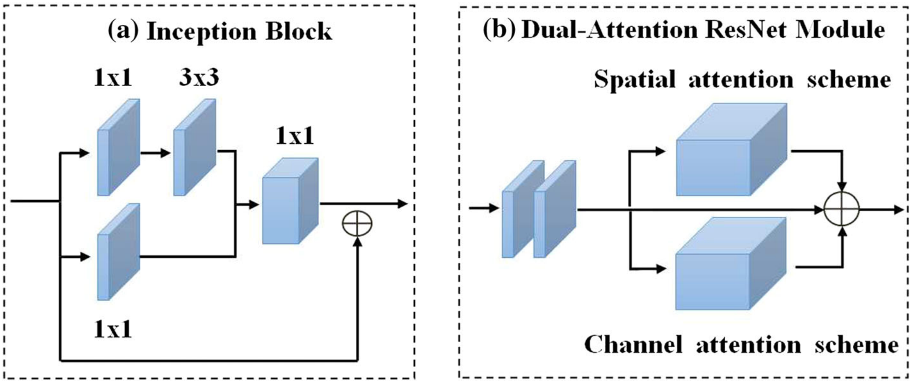

Fig. 3 Architecture of DRM, where **a** refers to inception block and **b** represents the dual combination form of DRM

which guarantees latter DRM structure could build highly convinced and effective context-wise descriptors.

3.3 Dual-Attention ResNet module

During the construction of deep neural networks, generated feature map contains different types of information such as low-frequency and high-frequency information across channels, patches and layers. All these information essentially have different reconstruction difficulties as well as different contributions to recover the implicit high-frequency details. However, most CNN-based methods consider different types of information are equal in informativeness and lack flexible modulation ability to deal with them. Moreover, simply increasing depth or width of network can hardly achieve better improvement with single-path direct connections or short skip connections among layers, since hierarchical features could hardly be fully utilized and long-term information that might be important for SR would be forgotten with the growing depth of network. Based on these two considerations, we construct a set of Dual-Attention ResNet modules (DRMs) and stack them in a chain structure to dynamically modulate multi-level features in a global-and-local manner. The proposed DRM structure emphasizes high informative and contribution information and suppresses redundant information, which guarantee to maintain long-term information and generate context feature map for image SR.

We design the proposed DRM to combine both channel and spatial attention schemes for generation of context-aware feature map, where its structure is shown in Fig. 3b. However, such complicated and multi-stage process of DRM make it hard to train for convergency. To ease this difficulty, we propose a residual shortcut on DRM, which makes gradient descent propagate in a much easier way.

How to appropriately combine channel and spatial attention schemes is discussed in many previous works. We show structures of these attention schemes in Fig. 4, including CBAM [47], CSAR block [15], RAM [20] and the proposed DRM. We could observe that DRM is different with comparative schemes in input feature map, designs of attention schemes and combination form. Without enough low-frequency features extracted from inputting LR image, it's hard to guarantee performance of constructed context-aware descriptors. We thus utilize a task-specified inception block before DRM structure for enhanced feature extraction.

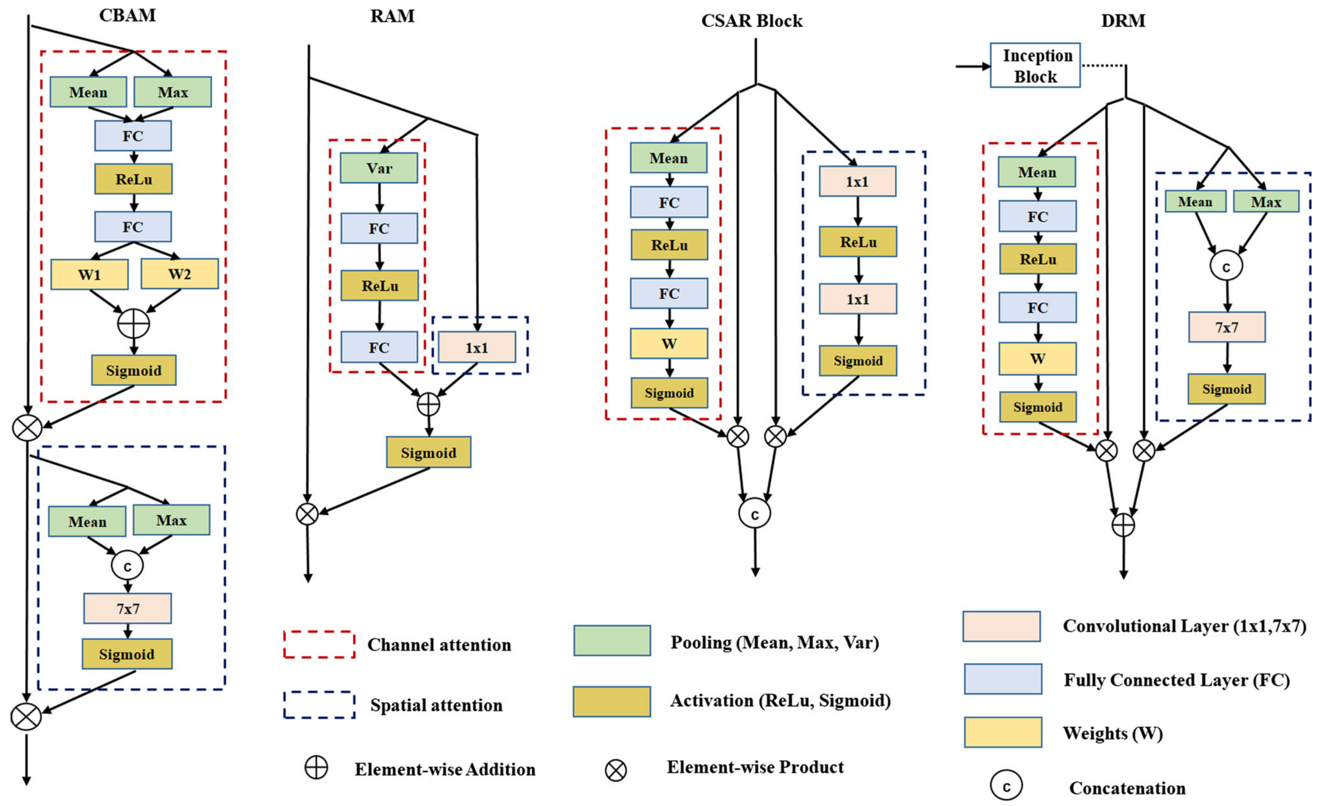

Fig. 4 Structure of various attention schemes, where CBAM can be regarded as a cascade form combination, DRM and CSAR represent dual form combination, and RAM is a fused form combination between two simplified attention schemes. It's noted that input feature map for DRM is previously enhanced by an inception block to increase its multi-level property

Applying an additional max-pooling operation to construct spatial attention scheme is adopted by CBAM and DRM, which exploits maximal context characteristics of feature map. Essentially, regions with maximal values could be edges, corners or places with high gradient values, which require more attention on their high-frequency details reconstruction. DRM thus utilizes max-pooling operation to detect such regions and adopts higher weights to emphasize these regions.

We prefer a dual form to combine the channel and spatial attention scheme, rather than a cascade form. We choose dual form for SISR on the basis of task purpose. For classification tasks, feature needs to be highly compressed to resolve high-level and semantic information. However, SISR task requires to restore high-frequency details based on feature maps. Generally speaking, cascade combination form leads passed-by information to be compressed, while dual form can increase bandwidth for information transmission to obtain abundant information for reconstruction. Last but not least, direct fuse combination form adopted by RAM works well only with simplified attention scheme, since fuse complicated multi-modal information is a challenging and unsolved problem. Therefore, a dual combination form is more favorable for SISR task.

We further improve dual combination form by replacing the last layer from concatenate operation to a simple element-wise addition operation, where the latter operation brings advantage of less parameters. Furthermore, back-propagation gradients can be equally passed to either channel or spatial attention scheme with element-wise addition operation, resulting in fast and stable convergency during training.

3.4 Structure of channel and spatial attention scheme

The proposed DRM aims to exploit inter-channel and intra-channel context relationship of feature maps with channel and spatial attention scheme, respectively. In this subsection, we describe structure details of the proposed channel and spatial attention schemes.

3.4.1 Channel attention scheme

A convolutional layer usually scans the input image and computes the corresponding 3D feature map. Considering that a convolutional layer consists of different channel filters, each 2D slice of the output 3D feature map essentially encodes spatial-visual responses raised by a channel filter. From the view of pattern recognition, each channel filter actually performs as a pattern detector. In other words, low-layer channel filters detect low-level visual cues like edges and corners, while high-layer ones detect high-level and semantic patterns like parts and objects [55]. By stacking different layers, a CNN extracts image features through a hierarchical representation of visual abstractions. Therefore, features extracted from CNN structure are essentially channel-wise and multilayer. However, not all the channel-wise features are equally important and informative for recovering high-frequency details. We thus utilize channel attention scheme to compute task-specified feature map for SISR by exploiting cross-channel relationship. In other words, channel attention scheme offers an intuitive descriptor on inherent context property among different feature channels.

As shown in Fig. 4b, global mean pooling is firstly performed on input feature map F_i to output global avg-pooled feature map F_c with size $C \times 1 \times 1$, which could be represented as

$$F_{\mathrm{c}} = AP(H_R(F_i)) \tag{11}$$

where function $H_R(\cdot)$ represents two Conv operations to generate features and is represented in Fig. 3b, and function $AP(\cdot)$ refers to the global mean pooling operation.

Then, F_{c} will be fed into a multilayer perception (MLP) with two hidden layers. It’s noted that the first hidden layer is used to perform a dimension reduction for a compact feature representation to aggregate information among channels. Finally, a sigmoid activation function is used to squeeze the output of MLP. Channel attention weight thus could be computed as:

$$C(F_i) = \mathrm{sig}(W_1 * (\mathrm{relu}(W_0 * F_{\mathrm{c}}))) \tag{12}$$

where functions $\mathrm{sig}(\cdot)$ and $\mathrm{relu}(\cdot)$ refer to sigmoid activation function and relu activation function, respectively, W_0 and W_1 are learnable parameter matrices and defined with size $\frac{C}{r} \times C$ and $C \times \frac{C}{r}$, respectively, and r is a pre-defined dimension reduction parameter and we set it as 16 by experiments.

3.4.2 Spatial attention scheme

We observe the information contained in feature maps or LR images is diverse over spatial positions. For example, the edge or texture regions usually contain high-frequency information, while the smooth areas have low-frequency information. To better recover high-frequency details and maintain low-frequency parts for SISR task, we thus propose a spatial attention scheme to adaptively optimize feature map in different regions with suitable operations. Moreover, human perception on visual effects requires high similarity between LR and HR images, which could be achieved by spatial attention scheme to globally adjust intensity distribution and visually enhance saliency regions. Spatial attention scheme is constructed based on difference of feature map of different positions, which essentially explores inter-channel relationship to construct context descriptor. Above all, spatial attention helps focus on saliency parts of feature map to describe inter-channel context property and is thus a beneficial complementary to channel attention.

As shown in Fig. 4b, a global max-pooling operation is first performed on input feature map F_i of DRM to output max-pooled feature map F_m with size $1 \times m \times n$, which could be represented as

$$F_m = MP(H_R(F_i)) \tag{13}$$

where function $H_R(\cdot)$ represents two Conv operations before attention scheme, and function $MP(\cdot)$ refers to the global max-pooling operation.

Then, we perform mean pooling operation along channel dimension to generate avg-pooled feature map F_a, which is different from F_c in channel attention scheme on its pooling direction. Afterward, a concatenation operation along channel axis is performed on avg-pooled feature map F_a and max-pooled feature map F_m. Finally, a convolutional layer with 7×7 kernel and a sigmoid function are performed on the concatenated feature map to generate spatial attention weight as follows:

$$S(F_i) = \mathrm{sig}(\mathrm{Conv}([F_\mathrm{a}, F_m]) \tag{14}$$

where [,] denotes concatenation operation along channel axis, function $\mathrm{Conv}(\cdot)$ means operation of a convolutional layer.

4 Experimental results

In this section, we show the effectiveness and efficiency of the proposed CASR network for SISR task. We first introduce dataset and measurements. Then, we conduct three groups of comparative studies to demonstrate CASR is effective in super-resolving real-world photos. Afterward, experiments on computational cost are conducted to prove the efficiency of CASR. Finally, we describe implementation details for readers' convenience.

4.1 Dataset and measurement

Among all popular dataset for super-resolution task, we choose five datasets including Set5 [2], Set14 [56], B100 [28], Urban100 [17] and Manga109 [10] for experiments, since Set5, Set14 and B100 consist of natural scene images, Urban100 contains challenging urban scenes images with quantity of visual details, and Manga109 is a dataset of Japanese cartoon drawing. It's noted that DIV2K [43] serving as a benchmark for NTIRE 2017 super-resolution Challenge, is adopted as part of training set. We achieve pairs of LR and HR images by bicubic operator on HR images. After such operations, we finally obtain 800 images for training and 100 images to perform cross-validation for all comparative SISR methods.

We choose two standard quality measures, i.e., peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM), to compare quality of super-resolution results. It's noted that PSNR is adopted to measure reconstruction quality by calculating power ratio of noisy signal introduced by SISR process. Higher PSNR value represents better quality of reconstructed image. Meanwhile, SSIM is used to quantify the similarities of structure between original and HR images. High SSIM value indicates that SISR doesn't affect basic structure of original image, thus proving good reconstruction quality.

Since the definition of PSNR is on the basis of MSE, we define all these three measures as follows:

$$\mathrm{MSE} = \frac{m * n}{\sum_{i=1}^{m}\sum_{j=1}^{n}(I(i,j) - P(i,j))^2} \tag{15}$$

$$\mathrm{PSNR} = 10 \times \log(\frac{255^2}{\mathrm{MSE}}) \tag{16}$$

$$\mathrm{SSIM}(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{17}$$

where m and n refer to the width and the height of the image, I and P represent output image after operation of super-resolution and the input original image, respectively, μ_x and μ_y represent the means of x and y, respectively, σ_x^2 and σ_y^2 represent the variances of x and y, respectively, σ_{xy} is the covariance of x and y, and c_1 and c_2 are two preset variables.

4.2 Performance and analysis

In this subsection, we conduct three groups of comparative experiments, where the first one is to check whether dual or cascade form is suitable for combination of channel and spatial attention schemes, the second group is designed to show the effectiveness of CASR with or without DRM, and the last one is performed on a variety of datasets to compare performance of CASR with other SISR methods. Before stating, we define number of DRMs utilized in CASR as n and filter number of each convolutional layer in CASR as k, which are key parameters during experiments.

We show statics of the first comparative experiment in Table 1, where we define $n = 16$ and $k = 64$. During experiments, we combine channel and spatial attention schemes in either cascade form or dual form, and keep other parts remain the same for a fair comparison. From Table 1, we can notice that PSNR/SSIM values achieved by CASR with dual form are higher than those with

Table 1 Comparisons on PSNR/SSIM measurement between cascade and dual combination form

Methods	Set5	Set14	B100	Urban100	Manga109	Average
Cascade	37.93/0.9603	33.43/0.9140	31.97/0.8971	31.78/0.9244	38.29/0.9762	34.68/0.9344
Dual	**37.97/0.9605**	**33.56/0.9163**	**32.06/0.8986**	**31.94/0.9261**	**38.39/0.9766**	**34.78/0.9356**

Bold values indicate the best performance among all comparative methods

The scale factor is set as ×2

Table 2 Comparisons on PSNR/SSIM measurements between CASR with or without DRM, where w and wo refer to with and without DRM, respectively, and the scale factor is set as ×2

Methods	Set5	Set14	B100	Urban100	Manga109	Average
CASR(wo)	37.85/0.9602	33.51/0.9159	32.06/**0.8989**	31.93/0.9265	38.05/0.9760	34.68/0.9355
CASR(w)	**37.94/0.9605**	**33.56/0.9165**	**32.07**/0.8987	**32.01/0.9267**	**38.45/0.9767**	**34.81/0.9358**

Bold values indicate the best performance among all comparative methods

cascade form on all the datasets. This proves that dual form is more appropriate than cascade form in DRM for context feature map generation. In fact, tasks, like recognition and detection, are in favor of cascade combination between different attention schemes, since comprehending high-level information or semantics meanings of images generally requires low-frequency features to be highly compressed for process of encoding and decoding. Meanwhile, dual form increases bandwidth of information transmission among modules of network, so that abundant information including both low- and high-frequency information, can be utilized by subsequent networks for image recovery or reconstruction tasks.

Details of the second comparative experiment are presented in Table 2, where we set $n = 16$ and $k = 64$. Specifically, we perform two rounds of experiments either with or without DRM structure represented as CASR(w) and CASR(wo), respectively. From Table 2, we can notice PSNR/SSIM values achieved by CASR(w) increase on the basis of CASR(wo) in most cases, which proves the effectiveness of DRM to involve context descriptors for feature map enhancement. For the case of B100 dataset where SSIM achieved by CASR(wo), i.e., 0.8989, is a bit larger than that obtained by CASR(w), i.e., 0.8987, we conclude that it sources from complicated and multiple categories of context information embedded in B100 dataset. With sufficient ability to encode context information of Manga109 dataset, CASR(w) obtains a much larger PSNR/SSIM value, i.e., 38.45/0.9767, than that of CASR(wo), i.e., 38.05/0.9760. In fact, Manga109 dataset consists of cartoon drawings with much more simple context information, compared with real-world natural images from B100 dataset.

Table 3 shows quantitative comparative results with various kinds of SISR algorithms for 2×, 3×, 4× and 8× SISR tasks, such as Bicubic, A+ [42], SRCNN [8], VDSR [18], EDSR [25], LapSRN [23], GuideAE [7], SRMDNF [57], IDN [61], MSLapSRN [22] and DualGAN [54]. It's claimed that we achieve results of all comparative methods directly from their published paper. With the same setting ($n = 32$ and $k = 128$), we could observe that CASR+ achieve better SISR performance than CASR due to its self-ensemble strategy, which pre-process input LR images by flip and rotation operations for data augmentation.

From Table 3, we could notice CASR+ achieves the best performance among Set5, Set14 and Manga109 datasets, since their context information with a few images or cartoon drawings could be encoded and described by DRM structure. Meanwhile, small increase or worse performance on PSNR/SSIM values can be viewed when comparing CASR+ with EDSR on B100 and Urban100, since images representing urban and natural scene are difficult and complex in modeling their context information. With only half number of filters compared with EDSR, CASR+ is able to produce superior reconstruction results in most cases, which proves DRM structure guarantee CASR+ to achieve good performance with a small amount of parameters. Essentially, involving context information by various attention schemes help focus on informative parts among different channels and regions to reduce computation burden and improve reconstruction performance. Compared with CASR+, G-GANISR [40] achieves better reconstruction results with 8× scale factor, which proves power of GAN-based structure for SR with larger scaling factor.

Samples of reconstruction visual effects are shown in Fig. 5 with 4× scale factor. We can notice that CASR+ accurately restores parallel straight lines and grid patterns like windows and building shape, since CASR+ guarantees to preserve low-frequency features. Blurry effects and loss of image details can be viewed in most cases achieved by comparative methods, since they fail to restore high-frequency details via training and learning on abundant low-frequency features. For example, we observe artifacts of

Table 3 Quantitative evaluation of state-of-the-art SISR algorithms, where average PSNR/SSIM for scale factors ×2, ×3, ×4, ×8 are listed, information not provided by original authors is marked with [–], and CASR+ represents application version of utilizing self-ensemble property for testing

Methods	Scale	Set5	Set14	B100	Urban100	Manga109
×2	Bicubic	33.66/0.9299	30.24/0.87688	29.56/0.8431	26.88/0.8403	30.80/0.9339
	A+ [42]	36.54/0.9544	32.28/0.9056	31.21/0.8863	29.20/0.8938	–/–
	SRCNN [8]	36.66/0.9542	32.45/0.9067	31.36/0.8879	29.50/0.8946	35.60/0.9663
	VDSR [18]	37.53/0.9590	33.05/0.9130	31.90/0.8960	30.77/0.9140	37.22/0.9750
	EDSR [25]	38.11/0.9602	33.92/0.9195	32.32/0.9013	**32.93/0.9351**	39.10/0.9773
	LapSRN [23]	37.52/0.9591	33.08/0.9130	31.08/0.8950	30.41/0.9101	37.27/0.9740
	GuideAE [7]	37.52/0.9591	33.08/0.9130	31.08/0.8950	30.41/0.9101	37.27/0.9740
	SRMDNF [57]	37.79/0.9601	33.32/0.9154	32.05/0.8984	31.33/0.9204	38.07/0.976
	IDN [61]	37.83/0.9600	33.30/0.9148	32.08/0.8985	31.27/0.9196	38.02/0.9749
	CASR	38.16/0.9612	33.86/0.9190	32.22/0.9007	32.70/0.9332	39.04/0.9779
	CASR+	**38.23/0.9614**	**34.02/0.9202**	**32.33/0.9016**	32.92/0.9347	**39.25/0.9784**
×3	Bicubic	30.39/0.8682	27.55/0.7742	27.21/0.7385	24.46/0.7349	26.95/0.8556
	A+ [42]	32.58/0.9088	29.13/0.8188	28.29/0.7835	26.03/0.7973	–/–
	SRCNN [8]	32.75/0.9090	29.30/0.8215	28.41/0.7863	26.24/0.7989	30.48/0.9117
	VDSR [18]	33.67/0.9210	29.78/0.8320	28.83/0.7990	27.14/0.8290	32.01/0.9340
	EDSR [25]	34.65/0.9280	30.52/0.8462	29.25/0.8093	28.80/0.8653	34.17/0.9476
	LapSRN [23]	33.82/0.9227	29.87/0.8320	28.82/0.7980	27.07/0.8280	32.21/0.9350
	GuideAE [7]	33.82/0.9227	29.87/0.8320	28.82/0.7980	27.07/0.8280	32.21/0.9350
	SRMDNF [57]	34.12/0.9254	30.04/0.8371	28.97/0.8025	27.57/0.8398	33.00/0.9403
	IDN [61]	34.11/0.9253	29.99/0.8354	28.95/0.8013	27.42/0.8359	32.69/0.9378
	CASR	34.67/0.9295	30.55/0.8448	29.15/0.8076	28.69/0.8625	34.06/0.9478
	CASR+	**34.75/0.9300**	**30.64/0.8467**	**29.29/0.8096**	**28.92/0.8660**	**34.42/0.9495**
×4	Bicubic	28.42/0.8104	26.00/0.7027	25.96/0.6675	23.14/0.6577	24.89/0.7866
	A+ [42]	30.28/0.8603	27.32/0.7491	26.82/0.7087	24.32/0.7183	–/–
	SRCNN [8]	30.48/0.8628	27.50/0.7513	26.90/0.7101	24.52/0.7221	27.58/0.8555
	VDSR [18]	31.35/0.8830	28.02/0.7680	27.29/0.0726	25.18/0.7540	28.83/0.8870
	EDSR [25]	32.46/0.8968	28.80/0.7876	27.71/**0.7420**	26.64/0.8033	31.02/0.9148
	LapSRN [23]	31.54/0.8850	28.19/0.7720	27.32/0.7270	25.21/0.7560	29.09/0.8900
	GuideAE [7]	31.54/0.8850	28.19/0.7720	27.32/0.7270	25.21/0.7560	29.09/0.8900
	SRMDNF [57]	31.96/0.8925	28.35/0.7772	27.49/0.7337	25.68/0.7731	30.09/0.9024
	IDN [61]	31.82/0.8903	28.25/0.7730	27.41/0.7297	25.41/0.7632	29.41/0.8936
	CASR	32.54/0.8995	28.79/0.7848	27.60/0.7383	26.49/0.7978	30.85/0.9133
	CASR+	**32.62/0.9000**	**28.89/0.7877**	**27.74**/0.7413	**26.73/0.8036**	**31.31/0.9175**
×8	Bicubic	24.40/0.6580	23.10/0.5660	23.67/0.5480	20.74/0.5160	21.47/0.6500
	SRCNN [8]	25.33/0.6900	23.76/0.5910	24.13/0.5660	21.29/0.5440	22.46/0.6950
	VDSR [18]	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
	MSLapSRN [22]	26.34/0.7558	24.57/0.6273	24.65/0.5895	22.06/0.5963	23.90/0.7564
	DualGAN [54]	–/–	–/–	27.85/0.8911	–/–	–/–
	G–GANISR [40]	**31.11/0.9082**	**28.07/0.8803**	**29.18/0.9065**	**27.23/0.8750**	–/–
	EDSR [25]	26.96/0.7762	24.91/0.6420	24.81/0.5985	22.51/0.6221	24.69/0.7481
	CASR	27.00/0.7755	24.98/0.6398	24.84/0.5974	22.49/0.6181	24.60/0.7799
	CASR+	27.24/0.7825	25.13/0.6439	24.90/0.5997	22.71/0.6256	24.88/0.7873

Bold values indicate the best performance among all comparative methods

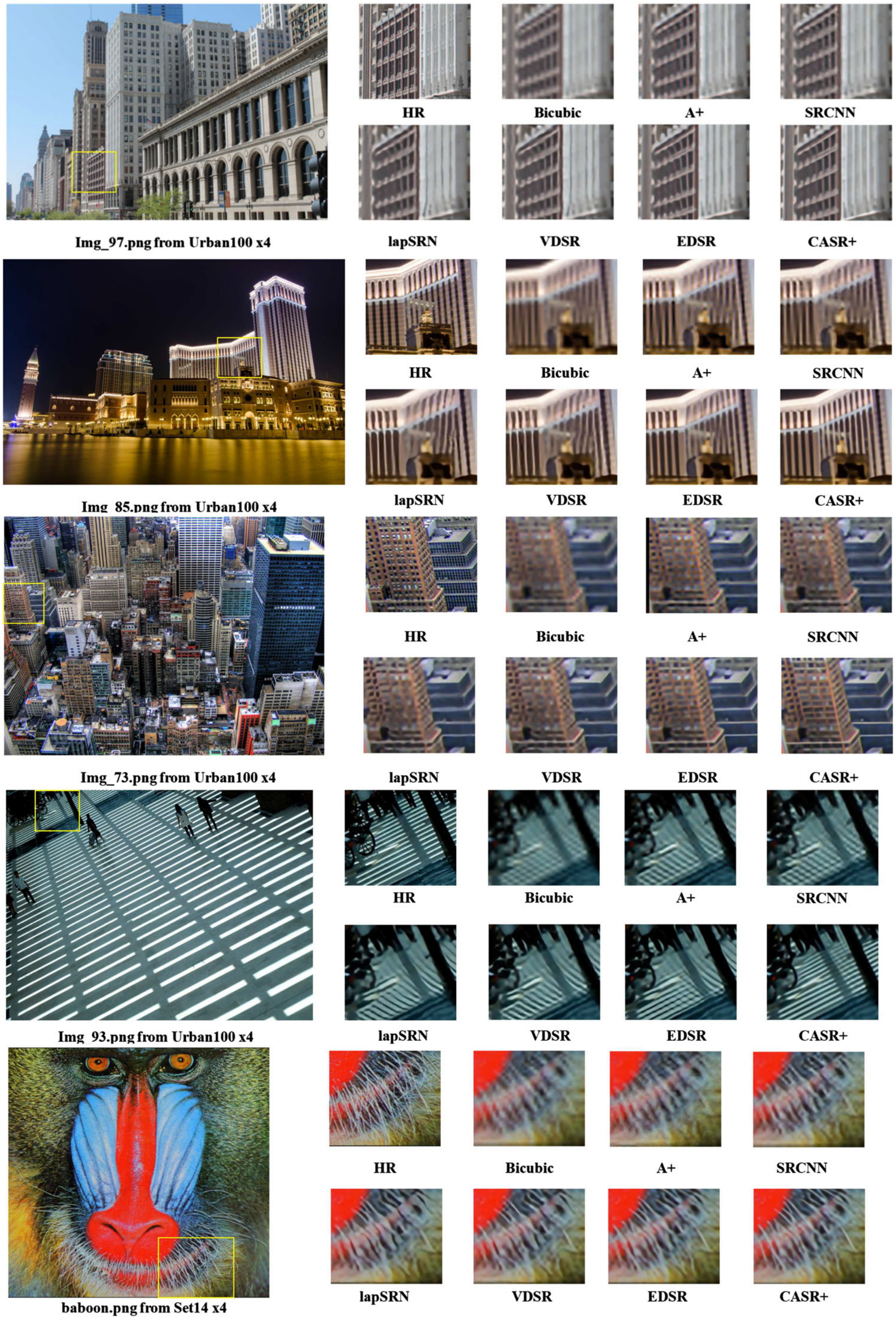

Fig. 5 Visual comparisons for 4× SISR on Set14 and Urban100 dataset, where yellow rectangle represents enlarge regions for comparisons

right slashes in reconstruction results of Img93 achieved by comparative methods. Meanwhile, our approach suppresses this artifact by constructing context descriptors for feature map enhancement, where we can clearly observe exact shape reconstruction as HR image achieved by CASR+.

Samples of reconstruction visual effects are shown in Fig. 6 with a larger scale factor, i.e., 8×. We could observe CASR+ achieves better reconstruction effects than comparative methods, even operating with a large scale factor. When compared with original HR images, we can notice some artifacts occur like unclear boundaries, blur effects and so on, which are caused by shortage of sufficient low-frequency information for reconstruction with large scale factor. We believe this situation can be improved by involving more training images, since we use a relatively small number of training images.

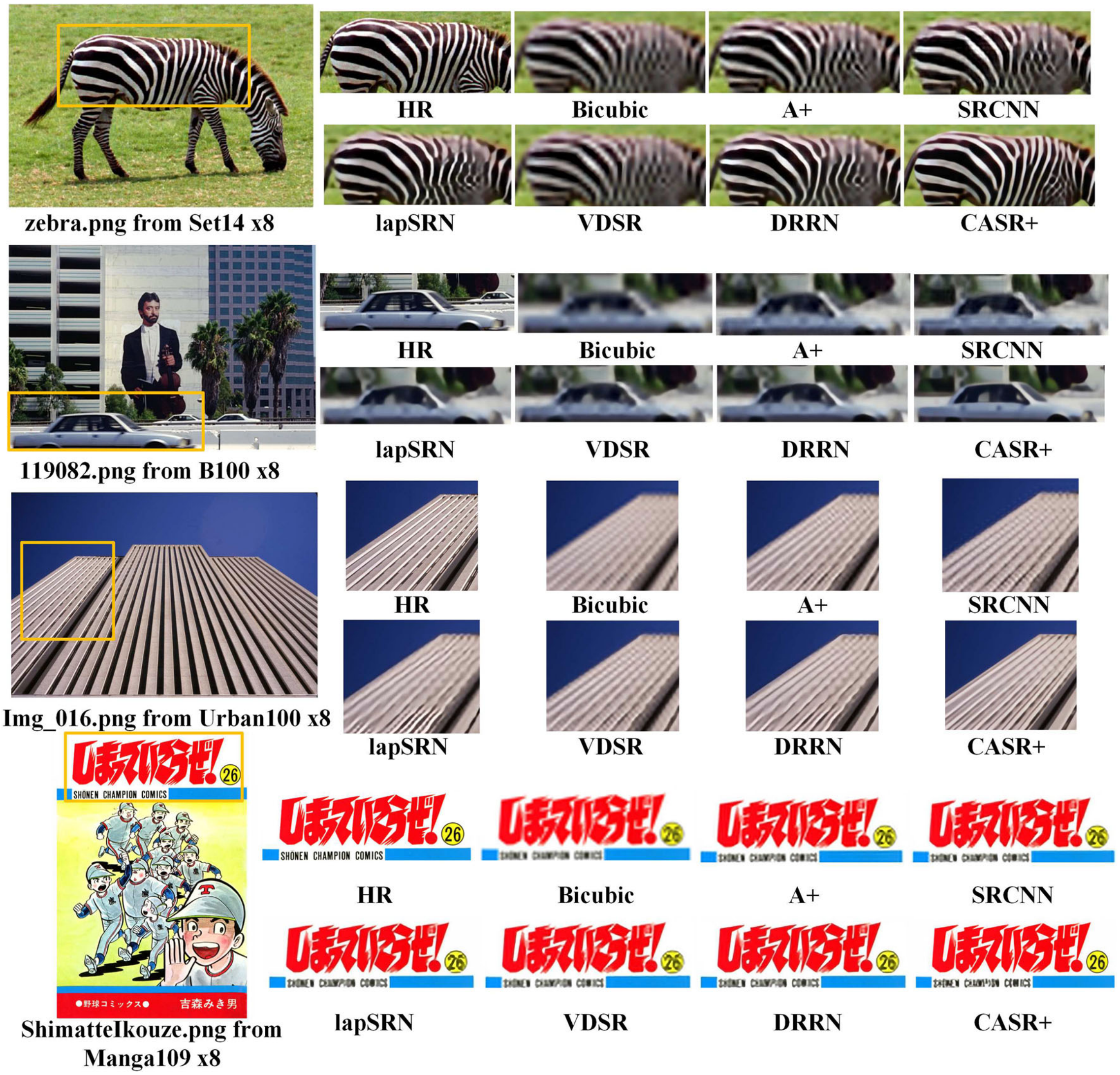

Fig. 6 Visual comparisons for 4× SISR on Set14, B100, Urban100 and Manga109 dataset, where yellow rectangle represents enlarge regions for comparisons

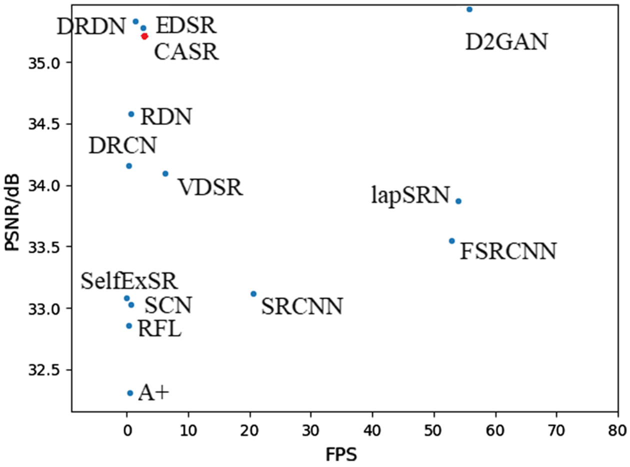

Fig. 7 Speed and accuracy trade-off, where runtime results are evaluated on five datasets with the scale factor 2×

4.3 Execution time analysis

We use original codes of comparative methods to evaluate runtime performance on the same machine with 3.4 GHz Intel i7 CPU (64G RAM) and NVIDIA Titan 1080Ti GPU (12G Memory). Figure 7 shows trade-offs between runtime and reconstruction performance (in terms of PSNR) on five datasets for 2× scale factor. Since CASR+ is computed by mean operation among 8 images for data augmentation, it's intuitive that CASR+ is almost 8 times lower than CASR, thus removing CASR+ from comparisons. From Fig. 7, we can notice running speed of CASR is a little faster than EDSR. Moreover, CASR achieves better performance on PSNR than EDSR. LapSRN [23], FSRCNN [9] and D2GAN [30] are fast enough to guarantee real-time performance; meanwhile, D2GAN is promising to achieve high-quality reconstruction results. SRCNN [8] is faster

Table 4 FPS evaluation between EDSR and CASR with different scaling factors

Methods	Scale	Set5	Set14	B100	Urban100	Manga109
EDSR [25]	×2	**3.35**	2.46	4.41	**1.25**	**1.54**
	×3	3.88	4.46	8.05	1.42	1.22
	×4	**8.42**	6.49	11.72	2.33	1.93
CASR	×2	3.27	**3.38**	**5.78**	1.07	0.93
	×3	**4.56**	**6.45**	**11.24**	**2.07**	**1.87**
	×4	8.23	**9.43**	**16.57**	**3.51**	**2.94**

Bold values indicate the best performance among all comparative methods

than CASR, DRDN [39], EDSR [25], DRCN [19], VDSR [18], SelfExSR [16], SCN [46], RFL [38], RDN [59] and A+ [42] achieve almost the same performance as CASR in runtime speed. However, all these methods get much lower PSNR values than CASR. Above all, CASR achieve a relatively good balance between runtime and reconstruction performance.

We show FPS performance comparisons with different scale factors in Table 4. We can observe CASR is much better in runtime performance than EDSR with large scaling factors, since context descriptor help reduce computation burden especially for cases of large scaling factors requiring more computing resources. Moreover, CASR can be deployed with 40MB storage size, which is one fourth of storage requirement by EDSR. We thus conclude that modeling context information by DRM not only helps CASR achieve better performance, but also leads to a lighter structure with smaller amount of parameters and less storage request.

4.4 Implementation Details

All of these experiments are performed on a single Titan 1080Ti GPU with 12GB memory. We set parameters of initial learning rate as 0.0001 and initial batch size as $48 * 48 * 3$. It's noted the learning rate is decremented by 0.5 for every 100 epochs and the total number of training epoch is 300. We adopt the Adam optimizer by setting its hyperparameters with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$. Our final CASR is trained within less than 30 h. In order to make full usages of training data, we used a data enhancement method, in which each training picture is rotated 90, 180, 270 degrees with a probability of 0.5, or flipped along a horizontal position.

5 Conclusion

In this work, we propose a lightweight context-aware residential network, named as CASR, which appropriately encodes channel and spatial attention information to construct context-aware feature map for SISR task. During construction, we propose an inception block to enhance feature representation, and a DRM to describe channel and spatial attention via dual form combination. During experiments, we conduct comparative experiments to test effectiveness of dual combination form, DRM attention structure, and CASR network. Compared with comparative methods, CASR achieves superior reconstruction performance and has advantages of less parameters, less memory request and faster running speed. With the development of cloud computing [26, 37], and mobile or wearable devices [11, 49], we believe a lightweight and real-time super-resolution method is required by applications. Therefore, our future work includes the explorations on improvements to achieve real-time performance and better visual effects with extreme imaging situations.

Acknowledgements This work was supported by National Key R&D Program of China under Grant 2018YFC0407901, the Natural Science Foundation of China under Grant 61702160 and 61602407, the Natural Science Foundation of Jiangsu Province under Grant BK20170892, Natural Science Foundation of Zhejiang Province under Grant LY19F030005 and LY18F020008, and the open Project of the National Key Lab for Novel Software Technology in NJU under Grant K-FKT2017B05.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Anderson P, He X, Buehler C, Teney D, Johnson M, Gould S, Zhang L (2018) Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of 2018 IEEE conference on computer vision and pattern recognition, pp 6077–6086
2. Bevilacqua M, Roumy A, Guillemot C, Alberi-Morel ML (2012) Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In: Proceedings of british machine vision conference
3. Bulat A, Yang J, Tzimiropoulos G (2018) To learn image super-resolution, use a gan to learn how to do image degradation first. In: Proceedings of European conference on computer vision, pp 185–200
4. Cao F, Li K (2018) A new method for image super-resolution with multi-channel constraints. Knowl Based Syst 146:118–128
5. Cao Q, Lin L, Shi Y, Liang X, Li G (2017) Attention-aware face hallucination via deep reinforcement learning. CoRR. arXiv:abs/1708.03132
6. Chen K, Yao L, Zhang D, Wang X, Chang X, Nie F (2019) A semisupervised recurrent convolutional attention model for human activity recognition. IEEE Trans Neural Netw Learn Syst. https://doi.org/10.1109/TNNLS.2019.2927224
7. Chen R, Qu Y, Li C, Zeng K, Xie Y, Li C (2019) Single-image super-resolution via joint statistical models-guided deep autoencoder network. Neural Computing and Applications pp 1–11

8. Dong C, Loy CC, He K, Tang X (2014) Learning a deep convolutional network for image super-resolution. In: Proceedings of European conference on computer vision, pp 184–199
9. Dong C, Loy CC, Tang X (2016) Accelerating the super-resolution convolutional neural network. In: Proceedings of European conference on computer vision. Springer, pp 391–407
10. Fujimoto A, Ogawa T, Yamamoto K, Matsui Y, Yamasaki T, Aizawa K (2016) Manga109 dataset and creation of metadata. In: Proceedings of the 1st international workshop on comics analysis, processing and understanding, p 2
11. Gong W, Qi L, Xu Y (2018) Privacy-aware multidimensional mobile service quality prediction and recommendation in distributed fog environment. Wireless Communications and Mobile Computing
12. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial nets. In: Proceedings of Neural Information Processing Systems, pp 2672–2680
13. Haris M, Shakhnarovich G, Ukita N (2018) Deep back-projection networks for super-resolution. In: Proceedings of computer vision and pattern recognition, pp 1664–1673
14. He T, Huang W, Qiao Y, Yao J (2016) Text-attentional convolutional neural network for scene text detection. IEEE Trans Image Process 25(6):2529–2541
15. Hu Y, Li J, Huang Y, Gao X (2018) Channel-wise and spatial feature modulation network for single image super-resolution. arXiv preprint arXiv:180911130
16. Huang J, Singh A, Ahuja N (2015) Single image super-resolution from transformed self-exemplars. In: Proceedings of IEEE conference on computer vision and pattern recognition, pp 5197–5206
17. Huang JB, Singh A, Ahuja N (2015) Single image super-resolution from transformed self-exemplars. In: Proceedings of computer vision and pattern recognition, pp 5197–5206
18. Kim J, Kwon Lee J, Mu Lee K (2016) Accurate image super-resolution using very deep convolutional networks. In: Proceedings of IEEE conference on computer vision and pattern recognition, pp 1646–1654
19. Kim J, Kwon Lee J, Mu Lee K (2016) Deeply-recursive convolutional network for image super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1637–1645
20. Kim JH, Choi JH, Cheon M, Lee JS (2018) Ram: Residual attention module for single image super-resolution. arXiv preprint arXiv:181112043
21. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Proceedings of neural information processing systems, pp 1097–1105
22. Lai W, Huang J, Ahuja N, Yang M (2017) Fast and accurate image super-resolution with deep Laplacian pyramid networks. CoRR abs/1710.01992
23. Lai WS, Huang JB, Ahuja N, Yang MH (2017) Deep laplacian pyramid networks for fast and accurate super-resolution. In: Proceedings of computer vision and pattern recognition
24. Ledig C, Theis L, Huszár F, Caballero J, Cunningham A, Acosta A, Aitken A, Tejani A, Totz J, Wang Z, et al (2017) Photo-realistic single image super-resolution using a generative adversarial network. arXiv preprint
25. Lim B, Son S, Kim H, Nah S, Lee KM (2017) Enhanced deep residual networks for single image super-resolution. In: Proceedings of computer vision and pattern recognition workshops, pp 1132–1140
26. Liu H, Kou H, Yan C, Qi L (2019) Link prediction in paper citation network to construct paper correlation graph. EURASIP J Wirel Commun Netw 1:233
27. Liu S, Huang D, Wang Y (2018) Receptive field block net for accurate and fast object detection. In: Proceedings of European conference on computer vision, pp 404–419
28. Martin D, Fowlkes C, Tal D, Malik J (2001) A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. Proc Int Conf Comput Vis 2:416–423
29. Mnih V, Heess N, Graves A, Kavukcuoglu K (2014) Recurrent models of visual attention. In: Proceedings of neural information processing systems, pp 2204–2212
30. Nguyen T, Le T, Vu H, Phung DQ (2017) Dual discriminator generative adversarial nets. In: Proceedings of Advances in neural information processing systems, pp 2670–2680
31. Qi L, Dou W, Chen J (2016) Weighted principal component analysis-based service selection method for multimedia services in cloud. Computing 98(1–2):195–214
32. Qi L, Xu X, Dou W, Yu J, Zhou Z, Zhang X (2016) Time-aware IoE service recommendation on sparse data. Mob Inf Sys 2016:4397061:1–4397061:12
33. Qi L, Dai P, Yu J, Zhou Z, Xu Y (2017) "time-location-frequency"-aware internet of things service selection based on historical records. Int J Distr Sens Netw 13(1):1–9
34. Qi L, Zhang X, Dou W, Ni Q (2017) A distributed locality-sensitive hashing-based approach for cloud service recommendation from multi-source data. IEEE J Sel Areas Commun 35(11):2616–2624
35. Qi L, Dou W, Wang W, Li G, Yu H, Wan S (2018) Dynamic mobile crowdsourcing selection for electricity load forecasting. IEEE Access 6:46926–46937
36. Qi L, Chen Y, Yuan Y, Fu S, Zhang X, Xu X (2019) A QoS-aware virtual machine scheduling method for energy conservation in cloud-based cyber-physical systems. World Wide Web. https://doi.org/10.1007/s11280-019-00684-y
37. Qi L, Wang R, Hu C, Li S, He Q, Xu X (2019) Time-aware distributed service recommendation with privacy-preservation. Inf Sci 480:354–364
38. Schulter S, Leistner C, Bischof H (2015) Fast and accurate image upscaling with super-resolution forests. In: Proceedings of IEEE conference on computer vision and pattern recognition, pp 3791–3799
39. Shamsolmoali P, Li X, Wang R (2019) Single image resolution enhancement by efficient dilated densely connected residual network. Signal Process Image Commun 79:13–23
40. Shamsolmoali P, Zareapoor M, Wang R, Jain DK, Yang J (2019) G-GANISR: gradual generative adversarial network for image super resolution. Neurocomputing 366:140–153
41. Tai Y, Yang J, Liu X (2017) Image super-resolution via deep recursive residual network. In: Proceedings of computer vision and pattern recognitio, pp 2790–2798
42. Timofte R, De Smet V, Van Gool L (2014) A+: Adjusted anchored neighborhood regression for fast super-resolution. In: Proceedings of Asian conference on computer vision. Springer, pp 111–126
43. Timofte R, Agustsson E, Van Gool L, Yang MH, Zhang L (2017) Ntire 2017 challenge on single image super-resolution: methods and results. In: Proceedings of computer vision and pattern recognition workshops, pp 114–125
44. Tong T, Li G, Liu X, Gao Q (2017) Image super-resolution using dense skip connections. In: Proceedings of international conference on computer vision, IEEE, pp 4809–4817
45. Wang Y, Perazzi F, McWilliams B, Sorkine-Hornung A, Sorkine-Hornung O, Schroers C (2018) A fully progressive approach to single-image super-resolution. In: Proceedings of IEEE conference on computer vision and pattern recognition workshops, pp 864–873

46. Wang Z, Liu D, Yang J, Han W, Huang TS (2015) Deep networks for image super-resolution with sparse prior. In: Proceedings of IEEE international conference on computer vision, pp 370–378
47. Woo S, Park J, Lee JY, So Kweon I (2018) Cbam: Convolutional block attention module. In: Proceedings of European conference on computer vision, pp 3–19
48. Xu X, Fu S, Qi L, Zhang X, Liu Q, He Q, Li S (2018) An IoT-oriented data placement method with privacy preservation in cloud environment. J Netw Comput Appl 124:148–157
49. Xu X, Li Y, Huang T, Xue Y, Peng K, Qi L, Dou W (2019) An energy-aware computation offloading method for smart edge computing in wireless metropolitan area networks. J Netw Comput Appl 133:75–85
50. Xu X, Liu Q, Luo Y, Peng K, Zhang X, Meng S, Qi L (2019) A computation offloading method over big data for iot-enabled cloud-edge computing. Future Gener Comput Syst 96:89–100
51. Xu X, Xue Y, Qi L, Yuan Y, Zhang X, Umer T, Wan S (2019) An edge computing-enabled computation offloading method with privacy preservation for internet of connected vehicles. Future Gener Comput Syst 95:522–533
52. Yan C, Cui X, Qi L, Xu X, Zhang X (2018) Privacy-aware data publishing and integration for collaborative service recommendation. IEEE Access 6:43021–43028
53. Yeung S, Russakovsky O, Jin N, Andriluka M, Mori G, Li F (2018) Every moment counts: Dense detailed labeling of actions in complex videos. Int J Comput Vis 126(2–4):375–389
54. Zareapoor M, Zhou H, Yang J (2019) Perceptual image quality using dual generative adversarial network. J Neural Comput Appl. https://doi.org/10.1007/s00521-019-04239-0
55. Zeiler MD, Fergus R (2014) Visualizing and understanding convolutional networks. In: Proceedings of European conference on computer vision, pp 818–833
56. Zeyde R, Elad M, Protter M (2010) On single image scale-up using sparse-representations. In: Proceedings of international conference on curves and surfaces. Springer, pp 711–730
57. Zhang K, Zuo W, Zhang L (2018) Learning a single convolutional super-resolution network for multiple degradations. In: Proceedings of IEEE conference on computer vision and pattern recognition, pp 3262–3271
58. 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 European conference on computer vision, pp 286–301
59. Zhang Y, Tian Y, Kong Y, Zhong B, Fu Y (2018) Residual dense network for image super-resolution. In: Proceedings of IEEE conference on computer vision and pattern recognition, pp 2472–2481
60. Zhao X, Sang L, Ding G, Han J, Di N, Yan C (2019) Recurrent attention model for pedestrian attribute recognition. In: Proceedings of the thirty-third AAAI conference on artificial intelligence, pp 9275–9282
61. Zheng H, Wang X, Gao X (2018) Fast and accurate single image super-resolution via information distillation network. In: Proceedings of IEEE conference on computer vision and pattern recognition, pp 723–731

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.