electronics

MDPI

Review

Attention Mechanisms in CNN-Based Single Image Super-Resolution: A Brief Review and a New Perspective

Hongyu Zhu [1], Chao Xie [1,2,*], Yeqi Fei [1,3] and Huanjie Tao [4]

[1] College of Mechanical and Electronic Engineering, Nanjing Forestry University, Nanjing 210037, China; czzhy100@yeah.net (H.Z.); feiyeqi@njfu.edu.cn (Y.F.)
[2] College of Landscape Architecture, Nanjing Forestry University, Nanjing 210037, China
[3] School of Intelligent Manufacturing, Nanjing University of Science and Technology Zijin College, Nanjing 210046, China
[4] School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China; huanjie_tao@nwpu.edu.cn
* Correspondence: chaoxie@njfu.edu.cn

Abstract: With the advance of deep learning, the performance of single image super-resolution (SR) has been notably improved by convolution neural network (CNN)-based methods. However, the increasing depth of CNNs makes them more difficult to train, which hinders the SR networks from achieving greater success. To overcome this, a wide range of related mechanisms has been introduced into the SR networks recently, with the aim of helping them converge more quickly and perform better. This has resulted in many research papers that incorporated a variety of attention mechanisms into the above SR baseline from different perspectives. Thus, this survey focuses on this topic and provides a review of these recently published works by grouping them into three major categories: channel attention, spatial attention, and non-local attention. For each of the groups in the taxonomy, the basic concepts are first explained, and then we delve deep into the detailed insights and contributions. Finally, we conclude this review by highlighting the bottlenecks of the current SR attention mechanisms, and propose a new perspective that can be viewed as a potential way to make a breakthrough.

Keywords: super-resolution; deep learning; convolution neural networks; attention mechanisms

Citation: Zhu, H.; Xie, C.; Fei, Y.; Tao, H. Attention Mechanisms in CNN-Based Single Image Super-Resolution: A Brief Review and a New Perspective. *Electronics* **2021**, *10*, 1187. https://doi.org/10.3390/electronics10101187

Academic Editor: Heidar Malki

Received: 20 April 2021
Accepted: 12 May 2021
Published: 15 May 2021

1. Introduction

Single image super-resolution, which will be abbreviated as SR hereinafter, aims at addressing the problem of reconstructing a high-resolution image from its given low-resolution counterpart, meanwhile refining the details and textures and improving the visual quality. SR is widely used in a variety of practical applications [1], ranging from remote sensing, video surveillance, medical imaging, to preconditioning of some high-level computer vision tasks, such as image classification and pattern recognition [2–4]. Currently, SR has been receiving an increasing amount of attention from both academic and industrial communities [5].

Although numerous methods have been proposed to push forward the SR performance and extensive research has been conducted to explore more effective and efficient ones, SR is still a fundamental and long-standing problem with many aspects to be promoted. This low-level vision task is challenging, since it is a severely ill-posed problem with a number of potential high-resolution images relating to the corresponding low-resolution one. Furthermore, with the increase in the scale factor, the ill-posed problem becomes more serious, and thus it needs more priors to determine the missing pixel values.

Recently, neural networks and deep learning have been widely used in the field of computer vision and pattern recognition [6]. Due to its inherent capability of overcoming

the drawbacks of traditional algorithms that rely heavily on hand-crafted features [7], deep learning gains great popularity and achieves tremendous success in the areas of computer vision [8,9], pattern recognition [10,11], speech recognition [12], etc. However, neural networks have faced some issues including their provability, stability, robustness, adversarial perturbations and noisy labels. Many researchers have noticed these problems and accordingly put forward their own views and solutions [13–16].

With the development of deep learning theory [17], convolutional neural networks (CNNs) [18–20] have attracted considerable attention from global researchers. Since Dong et al. [7,8] first proposed the pioneering work in SR, called SRCNN, it has been widely explored to design effective SR networks. Many studies have proved that deeper and wider SR networks generally achieve better results as compared to plain and shallow ones [21–24]. However, the growing parameters in deeper CNNs [25–27] also increase their training difficulty dramatically, making the networks harder to converge and optimize, which decreases the efficiency of each feature map in them. Fortunately, in recent years, a novel kind of technique called attention mechanisms [28–30], which was originally proposed to boost the representational power of deep networks, has been introduced to SR networks in order to help them perform better.

Although there are some existing CNN-based SR surveys in the literature [31–35], our work differs from them in that we concentrate on the networks that utilize attention mechanisms. Many existing surveys focus on the performance of SR [36], while ours pays more attention to the architecture of SR networks and their place of insertion of attention mechanisms.

As far as we are concerned, these attention mechanisms can be divided into three categories, and each category has an original form, then followed by its variants. The first class is called channel attention, which extracts the weight from each channel of the feature map to reweight itself. The second class, namely spatial attention, gains the weight matrix in 2D space for pixels at the same spatial position. The third class is non-local attention, which aims at calculating the weight of each position from the global perspective of a feature map.

In this survey, we briefly review these three kinds of mechanisms used in recent CNN-based SR and propose a new perspective to achieve further improvement. The rest of the paper is arranged as follows. In Section 2, the background of the SR, the CNN-based methods and the attention mechanisms used in the SR networks is presented. Section 3 gives the detailed explanation of the existing three attention mechanisms in super-resolution. In Section 4, we conclude the bottlenecks of the existing attention mechanisms used in SR networks and propose a new perspective, and Section 5 concludes the survey and discusses the future directions.

2. Background

In the task of SR, if we represent a high-resolution image by x and its low-resolution counterpart by y, the degradation process can be described by the following formula:

$$y = \phi\left(x, \theta_\eta\right) \tag{1}$$

where ϕ represents the degradation process, and θ_η represents the parameters, including the downscaling kernel and additive noise [37]. The SR solver tries to predict and reconstruct a high-resolution image counterpart $\hat{x}$ from the input low-resolution image y [38], which can be denoted as:

$$\hat{x} = \phi^{-1}\left(y, \theta_\varsigma\right) \tag{2}$$

where θ_ς is the parameters to make up the inverse-problem solver [39]. The complicated image degradation process is generally unknown and affected by various factors, such as

the occurrence of noise, blur, mosaic, compression, etc. In the research field, most of the researchers model the degradation as follows:

$$y = \left(x \otimes k\right) \downarrow_s + n \tag{3}$$

where $\otimes$ denotes the convolution operation, k denotes the convolution kernel which leads to blurring the images, and $\downarrow_s$ is the downscaling operation which reduces the height and width s times. n in the symbol represents the additive white Gaussian noise with kernel width σ, i.e., the noise level [40].

With a mushroom growth of deep learning technologies for the past several years, deep-learning-based SR models have been actively explored and have broken the previous SR performance record constantly. Various deep-learning-based methods are applied to the performance improvement, including CNN-based methods (e.g., SRCNN [7,8]), ResNet [41] based methods (e.g., VDSR [42] and EDSR [43]), and Generative Adversarial Nets (GAN) [44] based methods. Nevertheless, in this survey, we mainly focus on the attention-mechanism-based methods, which take advantages of various attention mechanisms to promote the effect of reconstruction. As mentioned previously, we divide this mechanism into three categories, each of which has a distinct characteristic and utilizes different dimensions of information from the feature map to reconstruct a more elaborate super-resolution image [45].

3. Attention Mechanisms in SR

3.1. Channel Attention Mechanism

In 2017, in order to boost the representational power and channel relationship, Hu et al. [28] proposed the SENet, which first develops the channel attention mechanism in order to fully use the different importance degree of different channels and mining the channel interdependence of the model. This mechanism is of great value for improving the efficiency of each feature map. The CNN based on the squeeze-and-excitation network leads to huge improvement in the classification networks, and it is widely used in designing neural networks for the down-streaming computer vision tasks [46].

In the image SR domain, researchers introduce this mechanism to the neural networks and thus promote the performance. RCAN [30] builds a CNN with the residual-skip-connection structure combined with the channel attention mechanism. SAN [47] refines the mechanism by using the covariance average pooling. DRLN [48] puts forward the mechanism, replacing the channel attention module with the proposed Laplacian module to learn features at multiple sub-band frequencies.

3.1.1. RCAN

The first channel-attention-based CNN model to solve the SISR problems was put forward by Zhang et al., namely very deep residual channel attention networks (RCAN) [30]. The proposed network has two contributions. The first contribution is the network structure RIR, which is the abbreviation of "Residual in Residual". The RIR structure, which is inspired by the famous architecture ResNet [41], contains the long skip connection (LSC), from behind the first residual group (RG) to after the last residual group, in order to pass the low-frequency information [49] from the front to the end, thus making it possible for the network to learn the residual information at a coarse level. The network accommodates 10 RGs. In each RG are 20 residual channel attention blocks (RCABs), and a short skip connection from behind the first RCAB towards after the end of the last RCAB. The two kinds of skip connections compose the RIR structure, which makes the network more stable to train.

The second highlight of the article, which is the main contribution, is the residual channel attention block (RCAB) that includes the channel attention operation. As shown in Figure 1a, each RCAB is composed of two convolution layers and one Rectified Linear Unit (ReLU) activation, followed by a channel attention unit. A skip connection connects

the front of the first convolution layer to the end of the channel attention block to pass forward residual features. In the channel attention block, a feature map with the shape H × W × C is then collapsed to the shape 1 × 1 × C, using the global average pooling operation, which computes the average value of each feature map. Then, a multilayer perceptron (MLP), which is called the gate mechanism, is used to mine the inside relation of the average value among each feature map channel. First, a convolution with kernel size 1 × 1 is utilized to shrink the shape to 1 × 1 × C/r, where r is the reduction ratio, the same as that in SENet. RCAN takes 16 as the ratio r. After a layer of ReLU activation, a 1 × 1 convolution is then exploited to upscale the size to the original 1 × 1 × C. After a sigmoid function, the weight of each channel has been completely generated. The newly generated weights have captured the relation and significance of each channel, so multiplying the weights with each corresponding channel, we obtain the final reweighted feature maps.

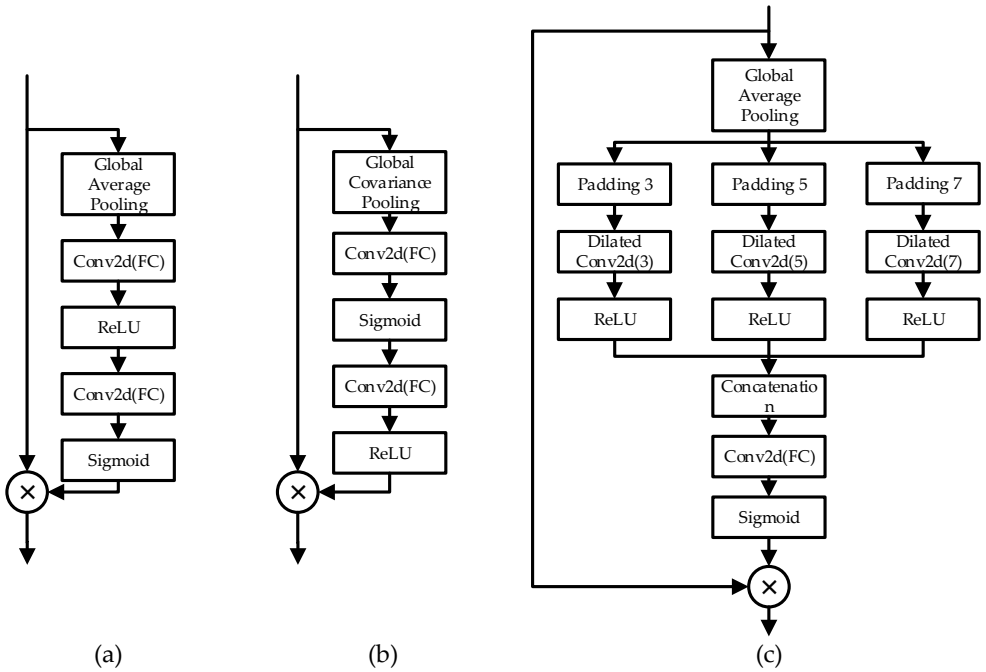

Figure 1. The detailed structure of channel attention mechanisms. (**a**) Channel attention in RCAN, (**b**) Second-order channel attention in SAN, (**c**) Laplacian pyramid attention in DRLN.

The introduction of the channel mechanism above significantly improves the performance and reduces the number of parameters required. The parameters of RCAN are about one-third of those of EDSR, which also has a Residual-in-Residual-like structure, but it achieves better performance.

RCAN uses L1 loss function and an ADAM [50] optimizer to train the network. The training data are 800 pairs of images in the DIV2K [51] dataset with data augmentation including rotating and flipping. In addition, compared to temporary methods like IRCNN [52] and VDSR [42], it brings better performance and certifies the positive effect of the channel attention mechanism.

3.1.2. SAN (Enhanced Channel Attention)

Dai et al. [47] proposed a Second-order Attention Neural network (SAN) for image super-resolution. It is pointed out that because the "global average pooling" manipulation, which is an import component in the channel attention in SENet [28] and RCAN [30], only explores the first-order statistics, while ignoring the statistics higher than the first-

order, this will result in the model's lack of discriminative ability. As second-order information has been proved helpful in large-scale visual recognition [53], the second-order channel attention mechanism is proposed to support the convolutional neural network.

Different from the first-order channel attention above, the second-order attention has more complicated steps, which are shown in Figure 1b. Denote a given feature map as $F = [f_1, f_2, \ldots, f_c]$ with the shape H × W × C, which has C channels and the size of H × W. First, reshape the feature map to a 2-D feature matrix X with the shape C × S, where S = H × W. Then, compute the sample covariance matrix using the following formula:

$$S = X\bar{I}X^T \tag{4}$$

where $\bar{I}$ is a matrix, with the value of all the diagonal elements set to $\frac{s-1}{s^2}$, and other elements set to $-\frac{1}{s}$. It can be computed by the formula $\hat{I} = \frac{1}{s}\left[I - \frac{1}{s}\mathbf{1}\right]$, where I is the s dimension identity matrix and **1** is a matrix of all ones.

The obtained matrix Σ, which is symmetric positive semi-definite, has the following eigenvalue decomposition (EIG):

$$\Sigma = U\Lambda U^T \tag{5}$$

where U denotes an orthogonal matrix while Λ is a diagonal matrix with eigenvalues $[\lambda_1, \lambda_2, \ldots, \lambda_C]$ in non-increasing order.

After the above-mentioned step is the covariance normalization operated on the obtained matrix Σ, which is equivalent to the power of eigenvalues:

$$\hat{Y} = \Sigma^{\alpha} = U\Lambda^{\alpha}U^T \tag{6}$$

The paper set $\alpha = \frac{1}{2}$ to achieve the best discriminative representations. Then, following SENet and RCAN, channel attention weight is computed. Denoting $\hat{Y}$ as $[y_1, y_2, \ldots, y_c]$, Dai et al. shrank it to the channel-wise statistics z with the shape 1 × 1 × C using global covariance pooling as the following formula:

$$z_c = H_{GCP}(y_c) = \frac{1}{C}\sum_{i}^{C} y_c(i) \tag{7}$$

After this operation, all the other steps are the same as in RCAN and SENet, which include MLP-like layers to fully exploit feature interdependencies from the aggregated information. Dai et al. also took r = 16 as the reduction ratio.

However, the second-order attention algorithm includes an EIG (eigenvalue decomposition) [54] method, which needs extra computational resource occupation, thus making it inefficient in training. The authors of SAN exploited the Newton–Schulz method [55] to solute the square root of the matrix Σ and accelerate the computation while training. First, pre-normalize the matrix Σ via the following equation:

$$\hat{\Sigma} = \frac{1}{tr(\Sigma)}\Sigma \tag{8}$$

where $tr(\Sigma)$ denotes the trace of Σ which is the sum of all the eigenvalues. Then, given $Y_0 = \hat{\Sigma}$, $Z_0 = I$, for n = 1, 2, …, N, the Newton–Schulz method [55] iterates the following equations alternatively:

$$Y_n = \frac{1}{2}Y_{n-1}\left(3I - Z_{n-1}Y_{n-1}\right) \tag{9}$$

$$Z_n = \frac{1}{2}\left(3I - Z_{n-1}Y_{n-1}\right)Z_{n-1} \tag{10}$$

After no more than five iterations, Y_N and Z_n quadratically converge to Y and Y^{-1}. Finally, there is the post compensation procedure, which can be expressed as:

$$\hat{Y} = \sqrt{tr(\Sigma)}Y_N \tag{11}$$

The main backbone of the SAN network, namely Non-locally Enhanced Residual Group (NLRG), consists of a Share-source Residual Group (SSRG) and two Region-level non-local modules (RL-NL) in the start and end of the network structure, which will be introduced in the next section. The Share-source Residual Group (SSRG) consists of 20 Local-source Residual Attention Groups (LSRAGs). Each LSRAG has 10 residual blocks and a second-order channel attention (SOCA) module behind them. SAN utilizes the L1 loss function and ADAM optimizer with 800 HR images in DIV2K to train the network. The SAN network achieved state-of-the-art results over other algorithms in the year of 2019.

3.1.3. DRLN

The Densely Residual Laplacian Network (DRLN) [48] for super-resolution by Anwar et al. introduced the Laplacian pyramid attention mechanism to the super-resolution domain, which is the most important insight of the creative work.

The major component of the DRLN network is the Cascading Residual on Residual module (CRIR), which has a long skip connection (LSC) to help the information flow through the cascading blocks. The CRIR architecture is mainly composed of cascading blocks, and each has a medium skip connection (MSC) to cascade feature concatenation. The cascading blocks are made of three dense residual Laplacian modules (DRLM) for each, and one DRLM consists of a densely connected residual unit [56], compression unit and Laplacian pyramid attention unit.

As shown in Figure 1c, the Laplacian pyramid attention module, which also computes the weight for each channel, has several differences against the channel attention module of RCAN. After the global average pooling operation to obtain a feature map with the size $1 \times 1 \times C$, which can be denoted as x, zero is used to pad x to the size of $7 \times 7 \times C$, $11 \times 11 \times C$, $15 \times 15 \times C$, denoted as c_1, c_2, c_3. Then, c_1, c_2, c_3 pass the dilated convolution layers with the kernel size 3 and dilated size 3, 5, 7, respectively.

The length of dilated convolution kernels just equals the size of the feature maps after padding. Same as RCAN, the reduction rate is set to 16 in the paper for each dilated convolution. After the three-pronged spear, the feature maps are concatenated and fed into a convolution layer to recover the dimension to the original $1 \times 1 \times C$. After a sigmoid function, the channel weights are generated to multiply each channel to obtain the final feature map.

The Laplacian pyramid attention mechanism has two main advantages over others suggested by the authors: first is its capability to learn features at multiple sub-band frequencies; second is its poser to adaptively rescale features and model feature dependencies at multiple feature spaces.

Same as RCAN, the network uses L1 loss function and an ADAM optimizer to help training. The well-designed architecture of the DRLN network takes the advantages of the residual connection, the dense concatenation, and the Laplacian attention to outperform to the classical network RCAN.

3.2. Spatial Attention Mechanism

3.2.1. SelNet

Choi and Kim et al. [57] proposed the super-resolution network SelNet with a novel selecting unit (SU). Different from the traditional ReLU activate function, which has the defect that it cannot back-propagate the training error through the switches while training the network, the proposed selecting unit works as a trainable switch. As shown in Figure 2a, SU consists of an identity mapping and a selection module (SM). The selection module is composed of a ReLU activation layer, a convolution layer with kernel size 1 × 1 and a sigmoid function layer in turn. The selection module computes the weight in the spatial domain, which can be regarded as belonging to the general spatial mechanism.

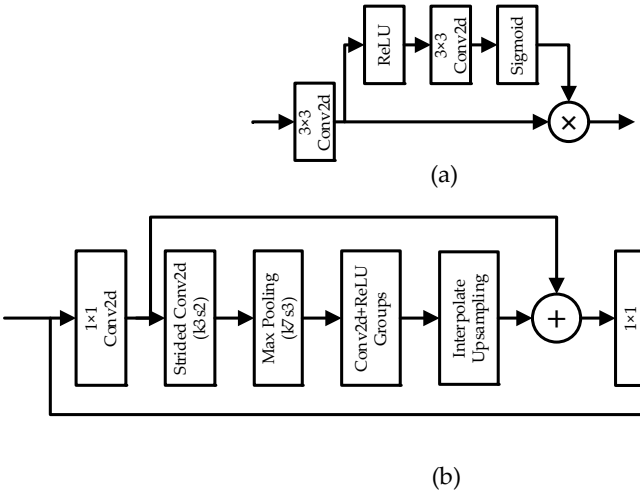

Figure 2. The detailed structure of spatial attention mechanism. (**a**) SU in SelNet, (**b**) EFA block in RFANet.

SelNet has 22 convolution layers in total, with one SU after each convolution layer. The authors adopted the enhanced residual connection to add the (n-2)-th feature map to the n-th feature map to feed forward to the next convolution layer. A sub-pixel layer is adopted to resize the feature map into the required height and width. Like VDSR, the input LR image is interpolated using the bicubic method and it adds the up-sized feature map to obtain the final SR image. All the manipulations are performed with the Y channel of the original image.

3.2.2. RFANet

Liu et al. [58] proposed the Residual Feature Aggregation Network for Image Super-Resolution (RFANet), which enhances the spatial attention to make a better improvement.

The main architecture of the network is the 30 residual feature aggregation (RFA) modules with a residual skip connection. The RFA module contains four residual blocks and a 1 × 1 convolutional layer. The residual features of the first three blocks are sent directly to the end of the RFA module and concatenated together with the output of the fourth residual block. This creative way makes the fullest use of all these residual features. The RFA module includes a convolution layer, a ReLU activation layer, a convolution layer, and an ESA block, which utilizes the enhanced spatial attention to promote the performance.

The ESA block, which is shown in Figure 2b, starts with a 1 × 1 convolution layer, which can decrease channel dimensions so as to lightweight the network. Then, a convolution with stride = 2 is utilized to shrink the height and width of the feature map, which is followed by a max-pooling layer with a large receptive field with a 7 × 7 kernel and

taking 3 as stride. An up-sampling layer is then added, which uses bilinear interpolation as the strategy to recover the feature map to the original height and width. A skip connection is built from the reduced-channel feature map to after the up-sampling layer in the end. Finally, a 1 × 1 convolution layer helps to restore the number of channels, followed by a sigmoid function layer to generate the attention mask. Multiplying the mask and the feature map, we can obtain the reweighted value.

The RFANet uses the L1 loss function and ADAM optimizer to help with the training. A lot of the ablation studies and experiments to combine the RFA blocks with other baselines prove the effect of the proposed method.

3.3. Combining the Above Two Attention Mechanisms

3.3.1. CSFM

Hu et al. proposed the Channel-wise and Spatial Feature Modulation Network for Single Image Super-Resolution (CSFM) [59], which combines the channel attention and the spatial attention mechanism to take the advantages of both of them.

The proposed CSFM network, which is shown in Figure 3a, consists of three blocks: an initial feature extraction sub-network (IFENet), a feature transformation sub-network (FTNet) and an upscaling sub-network (UpNet). The FTNet, which is the main part of the network, is composed of eight feature modulation memory modules (FMM) as a building module and stacks several FMM modules within a densely connected structure. An FMM module contains a channel-wise and spatial attention residual (CSAR) blockchain and a gated fusion (GF) node. The CSAR blockchain has 20 CSAR blocks.

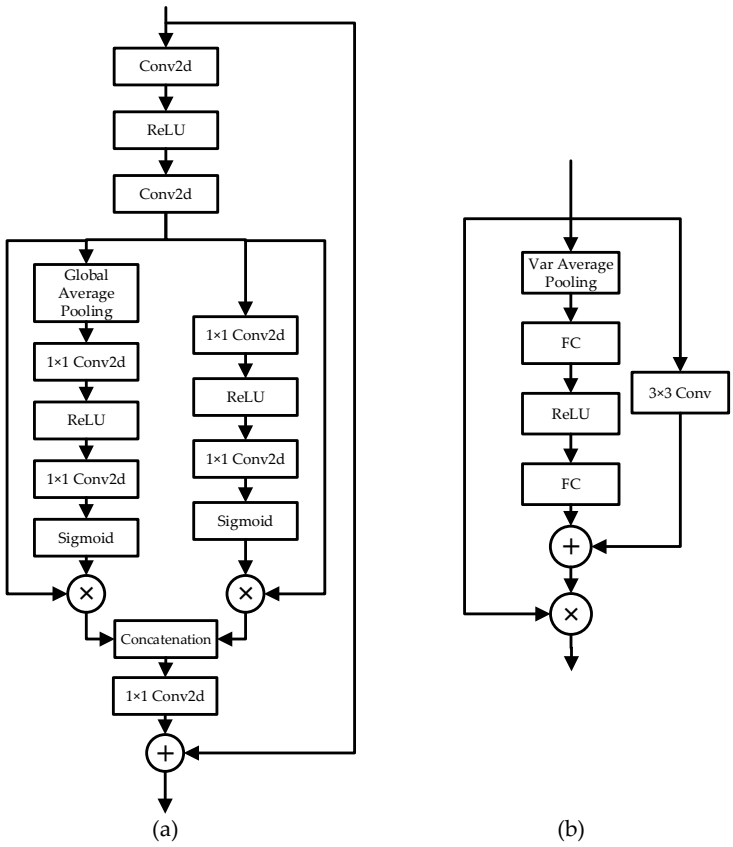

Figure 3. The detailed structure of the attention blocks, which has the combination of the two kinds of attention mechanisms. (**a**) CSAR module in CSFM, (**b**) RAM in SRRAM.

In a CSAR block, in order to increase the perception of features with a higher contribution and value, thus increasing the discriminating capability of the model, the authors designed a structure that takes channel-wise attention (CA) and spatial attention (SA) in a parallel position and combines them together.

The CA unit works the same as the channel mechanism in RCAN, and has the same reduction ratio of 16. The SA unit is utilized concurrently to learn the diverse information of the feature map in the spatial domain and enable the network to discriminate the importance of different regions. Let $U = [u_1, u_2, \ldots, u_c]$ be an input to the SA unit, which has the shape of C × H × W, and the input will pass two convolution layers. The first one increases the number of channels two-fold, then the next convolution layer reduces the number of channels to 1, i.e., the shape of the feature map has been changed to 1 × H × W. After a sigmoid function layer to generate the SA mask, multiply the original feature map with the mask and the output is generated.

Different from the enhanced spatial attention (ESA) block in RFANet, the SA in CSAR only generates a 1 × H × W mask while the mask in ESA has the same number of channels with the feature map, i.e., every value in the feature map has its weight value to multiply with.

After the parallel CA and SA unit is the concatenating manipulation to connect the output of both of them, and a 1 × 1 convolution layer is used to adaptively fuse two types of attentive features with learned weights.

Another insight of CSFM is its gate node, which is designed to integrate the information coming from the previous FMM modules and from the current blockchain through an adaptive learning process. The network uses the L1 loss function for training and is optimized by the ADAM optimizer. All the proposed mechanisms of CSFM have been proved effective through the ablation study.

3.3.2. SRRAM

Kim and Choi et al. [60] proposed the Residual Attention Module (RAM, shown in Figure 3b) for Single Image Super-Resolution. The RAM module combines the channel attention and the spatial attention, but with some changes that better fit the super-resolution task.

In the Residual Attention Module, after a convolution, a ReLU activation and a convolution layer, the feature maps are sent into the channel attention (CA) unit and the spatial attention (SA) unit, respectively. Different from the previous works, the authors proposed that, since SR ultimately aims at restoring high-frequency components of images, it is more reasonable for attention maps to be determined using high-frequency statistics about the channels, so the variance pooling methods are adopted instead of the global average pooling. The rest of the CA is the same as that in RCAN. For the spatial attention (SA) unit, it is claimed that each channel represents a kind of filter and different filters are used to extract different features. It is of great importance to deal with each channel on its merits. Depth wise convolution with kernel size as 3 × 3 is chosen to generate the SA mask. Finally, the mask of CA and SA is added and passes a sigmoid activation function to be the final mask of the RAM module.

SRRAM has an RCAN-like architecture and has 16 RAM residual blocks. The joining of the proposed mechanism is proved extremely successful.

3.4. Non-Local Attention

Wang et al. [61] first proposed the Non-local Neural Network, which is further researched and used in building deep-learning networks for low-level tasks such as super-resolution. Zhang et al. [62] proposed Residual Non-local Attention Networks for Image Restoration, which employ a pixel-level non-local attention mechanism to boost the performance for low-level vision tasks such as super-resolution and denoising. As shown in Figure 4a, given image feature map X, the non-local attention is defined as:

$$Z_{i,j} = \sum_{g,h} \frac{\exp\left(\phi\left(X_{i,j}, X_{g,h}\right)\right)}{\sum_{u,v} \exp\left(\phi\left(X_{i,j}, X_{u,v}\right)\right)} \psi\left(X_{g,h}\right) \tag{12}$$

where (i, j), (g, h) and (u, v) are pairs of coordinates of X. $\psi(\cdot)$ is the feature transformation function, and $\phi(\cdot, \cdot)$ is the correlation function to measure similarity that is defined as:

$$\phi\left(X_{i,j}, X_{g,h}\right) = \theta\left(X_{i,j}\right)^T \delta\left(X_{g,h}\right) \tag{13}$$

where $\theta(\cdot)$ and $\delta(\cdot)$ are feature transformations. Note that the pixel-wise correlation is measured in the same scale. The SAN network utilizes the region-level non-local attention block. The CSNLN uses the enhanced non-local attention, which extracts cross-scale features to help reconstruct the HR images. The PAN further uses the non-local information among different scales with pyramid-shaped feature maps.

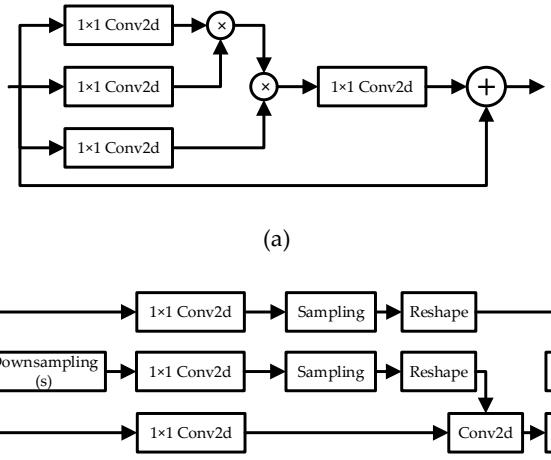

Figure 4. The detailed structure of non-local attention. (**a**) Non-local attention, (**b**) Cross-scale non-local attention.

3.4.1. SAN (Region-Level Non-Local Attention)

Except for the second-order attention mechanism, which is the greatest contribution of SAN, it also first employs the Region-Level Non-Local (RL-NL) to further capture the correlation of long-range dependencies throughout the entire feature map, as the low-level vision tasks prefer the non-local operations at a proper neighborhood size, which is proved in [62]. Therefore, the feature map is divided into a grid of regions, which is 2 × 2 in the paper. The four patches compute their non-local maps and values, respectively, and then are merged to compose a whole feature map. Additionally, this manipulation reduces the computing burden.

The RL-NL modules are placed at the beginning and end of the backbone network to exploit the spatial correlations of features.

3.4.2. CSNLN

Recently, Mei et al. [63] proposed an innovative design of a network which can super-resolve images with Cross-Scale Non-Local (CS-NL) attention and exhaustive Self-Exemplars Mining (SEM), which enables the network to fully excavate the self-similarity [63].

The main architecture of the designed network is composed of the recurrent of the Self-Exemplars Mining (SEM) cell, which fuses the features from the cross-scale non-local

attention module, the in-scale non-local attention module and the local branch. All the fused feature maps from SEM cells are concatenated at the end of the main structure to reconstruct high-resolution images.

The in-scale non-local attention module has the same operation as in [47], except for the deconvolution layer at the end position to upscale the feature map to match the output of the cross-scale non-local attention module. The cross-scale non-local attention module in Figure 4b, which is newly proposed by the authors, is designed to measure the correlation between low-resolution pixels and larger-scale patches in the LR images. Denote the size of the feature map input as W × H. One of the branches of the module down-samples the feature map, using the bilinear method, followed by a reshape operation to transform the feature map to the filters with kernel size p. Another branch has a 1 × 1 convolution to reduce the channels to generate the input and uses the filters generated to perform the convolution. After a softmax layer, the weighted tensor is prepared for deconvolution. The last branch uses convolution and the reshape operation to transform the feature map to the filters with kernel size *s*p × *s*p to be the filters for deconvolution. The final operation is the deconvolution, and we obtain the feature map upscaled with the shape *s*W × *s*H.

With three feature maps from the IS-NL module, the CS-NL module and the local branch, instead of concatenating them together, the authors creatively proposed a mutual-projected fusion to progressively combine features together. The residual between the CS-NL feature and the IS-NL feature after convolution is added to the IS-NL feature map, and the residual between the added result and the local feature is also added to calculate the final feature.

3.4.3. Pyramid Attention Networks

Mei et al. [64] recently proposed a new Pyramid Attention Network (PANet), which is the updated version of the cross-scale non-local network. The pyramid attention is a "plug and play" module. They achieved state-of-the-art performance in several image restoration tasks, especially when plugged into the EDSR network, and it outperformed the original EDSR and the SAN network, which was the best model in 2019.

It is pointed out that recent non-local-attention-based methods show restricted performance due to the simple single-scale correlations, further reduced by involving many ill matches during the pixel-wise feature matching in attention units. In order to make good use of the correlation of image patches in different sizes, a new type of non-local attention module is proposed, which downscales the feature map to the shape of a pyramid, namely the pyramid attention module.

Unlike the original attention module, which can be presented briefly:

$$y^i = \frac{1}{\sigma(x)} \sum_j \phi\left(x^i, x^j\right) \theta\left(x^j\right) \tag{14}$$

where *i*, *j* are indices on the input *x* and output *y*, respectively, the Scale Agnostic Attention, which composes the pyramid attention module, can be expressed as:

$$y^i = \frac{1}{\sigma(x)} \sum_{s \in S} \sum_j \phi\left(x^i, x^j_{\delta(s)}\right) \theta\left(x^j_{\delta(s)}\right) \tag{15}$$

where $S = \{1, s_1, s_2, \ldots, s_n\}$ is a series of given factors to downscale the feature map, and the $\delta(s)$ represents a s^2 neighborhood centered at index *j* on input *x*. The function ϕ computes pair-wise affinity between two input features. θ is a feature transformation function that generates a new representation of x_j. The output response y_i obtains information from all features by explicitly summing over all positions and is normalized by a scalar function $\sigma(x)$.

Built by the Scale Agnostic Attention, the pyramid attention block has the following expression:

$$y^i = \frac{1}{\sigma(x, \mathcal{F})} \sum_{z \in \mathcal{F}} \sum_{j \in z} \phi\left(x^i, z^j\right) \theta\left(z^j\right) \tag{16}$$

where $F = \{F_1, F_2, \ldots, F_n\}$ are generated with the scale factor series S, i.e., F_i has the shape $\frac{H}{s_i} \times \frac{W}{s_i}$.

In the article, the pair-wise function is determined to be the embedded Gaussian function, which has the following formula:

$$\phi\left(x^i, z^j\right) = e^{f\left(x^i\right)^T g\left(z^j\right)} \tag{17}$$

where $f(x^i) = W_f x^i$ and $g(z^i) = W_g z^j$. A simple linear function $\theta = W_\theta z^j$ is chosen to be the feature transformation and scalar function $\sigma(x, F) = \sum_{z \in F} \sum_{j \in z} \phi(x^i, z^j)$ is set.

When adding the module into the EDSR network, S, which is the collection of scale factor S, is set to {1.0, 0.9, 0.8, 0.7, 0.6}, leading to the 5-layer feature pyramid within the attention block. The attention block is plugged into the middle position of the backbone of the EDSR network, namely PA-EDSR. It is proposed that the PA-EDSR reconstructs more accurate image details for its excellent mechanism to utilize the features across the whole image from distinct scales.

4. Bottlenecks in SR Attention Mechanisms and a New Perspective

Although the attention mechanisms mentioned above have been proven helpful in the SR field, they still have some shortcomings. The traditional channel attention in SE and RCAN, which briefly squeezes the 2D feature map by coarsely doing the global average pooling (GAP) in order to generate the weight for each channel, only considers re-weighing each channel by modeling channel relationships, thus ignoring the coordinate of the features. The spatial attention mechanism and non-local attention mechanism, which have been proved effective when joint to the super-resolution network, need a large sum of calculation, which increases the amount of the parameters and prolongs the inferring time.

We are also exploring an enhanced mechanism, which ought to have a plug-and-play attribute and a light-weighted structure. When plugged into the SR networks, it should greatly improve their performance.

The coordinate attention, which was proposed recently by Hou et al. [65], has gained increasing notice due to its excellent performance when added into the neural networks for image classification and its down-streaming tasks such as segmentation and object detection [66,67]. Figure 5 shows the structure of this creative attention mechanism. As a new branch of the channel attention, to effectively solve the problem, the newly proposed method first calculates the mean value in each channel of the feature map in the x- and y-coordinate, that is, using the two spatial extents of pooling kernels (H, 1) and (1, W) to encode each channel along the horizontal coordinate and the vertical coordinate, respectively, formulated by the following equations:

$$z_c^h(h) = \frac{1}{W} \sum_{0 \le i < W} x_c(h, i) \tag{18}$$

$$z_c^w(w) = \frac{1}{H} \sum_{0 \le j < H} x_c(j, w) \tag{19}$$

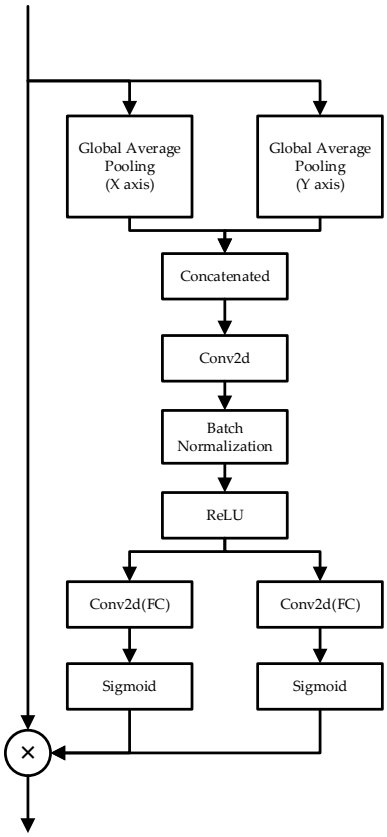

Figure 5. The detailed structure of coordinate attention mechanism.

The two formulas aggregate the information in the height and width direction, yielding a pair of direction-aware feature maps. The two feature maps, which contain the positional information, have larger capacity than only using the global average pooling. The proposed transformation also allows the next steps to capture the long-range dependencies along one spatial direction and preserve precise positional information along the other spatial direction, thus helping the networks locate the features precisely while training and inferring.

Then, the two groups of the obtained arrays are concatenated to pass a convolution layer with kernel size 1 × 1, which reduces the number of channels, where the transformation expression can be written as:

$$f = \delta\left(F_1\left(\left[z^h, z^w\right]\right)\right) \tag{20}$$

where $[\cdot, \cdot]$ denotes the concatenation operation along the spatial dimension, δ is a nonlinear activation function and $f \in R^{C/r\times(H+W)}$ is the intermediate feature map that encodes spatial information in both the horizontal direction and the vertical direction. r is the reduction ratio for controlling the block size as in the SE block.

Then, the feature map is split into the original two groups according to the original proportion, which can be denoted as $f^h \in R^{C/r\times H}$ and $f^w \in R^{C/r\times W}$.

Next, each group performs a convolution operation to restore the number of channels. Finally, after a sigmoid activation, the weight in the x- and y- coordinate is generated to reweigh the raw feature map in the two directions, whose process can be denoted as:

$$g^h = \sigma\left(F_h\left(f^h\right)\right) \tag{21}$$

$$g^w = \sigma\left(F_w\left(f^w\right)\right) \tag{22}$$

Thus, the calculation of the final feature map is written as:

$$y_c(i,j) = x_c(i,j) \times g_c^h(i) \times g_c^w(j) \tag{23}$$

In the super-resolution field, the location information holds tremendous importance because the reconstruction of the image needs the pixel value of each position precisely. Additionally, different positions in the feature maps have different significance when contributing to the reconstructed super-resolved images.

5. Conclusions

Attention mechanisms have been proved a very helpful method to help enhance the performance of convolutional neural networks for image SR. As the research of deep-learning and super-resolution continues, many new mechanisms are proposed, which can be classified into three kinds: channel attention, spatial attention and non-local attention mechanisms. We have a comprehensive survey over these methods, introducing the detailed principles and steps of them and their variants, with the accurate architecture information of the particular neural networks. In Section 3.1, RCAN, SAN, and DRLN are explicitly introduced, which contain the raw and variant versions of channel attention mechanisms, and all gain great improvement beyond the baseline without attention mechanisms. In Section 3.2, we introduce the spatial attention mechanism, which consists of the SelNet and RFANet. They utilize the spatial information inside the feature map to help reconstruct better high-resolution images. CSFM and SRRAM are featured in Section 3.3, which both have the combination of the two mechanisms. Non-local attention is exhibited in Section 3.4, including the SAN, CSNLN, and PA-EDSR networks. They explore the global correlation in the feature map, thus they perform well when there are similar patents and features in the images. We show precise analyses of their advantages and shortcomings. The performance of each network with its particular attention mechanism is shown in Table 1. Furthermore, we introduce a new perspective, namely the coordinate attention, which belongs to the channel attention mechanism but avoids the problem of the neglect of the position information in the primitive channel attention mechanism. With its distinct operating process, the newly proposed method is certain to surpass the former networks when integrated into a well-designed network structure. The proposed mechanism can also be plugged into CNNs for other tasks in order to push forward their performance.

Table 1. The scores of different SR network on Set5 [68] and sources of all networks mentioned in the paper. The EDSR network is set to be the baseline without any attention mechanism. The best scores are in bold.

SR Networks	x2		x3		x4		x8		Attention Mechanisms			Sources
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	CA	SA	NLA	
EDSR [43]	38.11	0.9601	34.65	0.9282	32.46	0.8968	-	-				CVPR2017(BASELINE)
RCAN [30]	38.27	0.9614	34.74	0.9299	32.63	0.9002	27.31	0.7878	√			ECCV2018
SAN [47]	38.31	**0.9620**	34.75	0.9300	32.64	0.9003	27.22	0.7829	√		√	CVPR2019
DRNL [48]	38.27	0.9616	34.78	0.9303	32.63	0.9002	**27.36**	**0.7882**	√			TPAMI2020(Arxiv2019)
SelNet [57]	37.98	0.9598	34.27	0.9257	32.00	0.8931	-	-		√		CVPRW2017
RFANet [58]	38.26	0.9615	34.79	0.9300	**32.66**	0.9004	-	-		√		CVPR2020
CSFM [59]	38.26	0.9615	34.76	0.9301	32.61	0.9000	-	-	√	√		TCSVT2018

SRRAM [60]	37.82	0.9592	34.30	0.9256	32.13	0.8932	-	-	√	√		Neurocomputing2020(Arxiv2018)
CSNLN [63]	38.28	0.9616	34.74	0.9300	32.68	0.9004	-	-			√	CVPR2020
PA-EDSR [64]	**38.33**	0.9617	**34.84**	**0.9306**	32.65	**0.9006**	-	-			√	Arxiv2020

Author Contributions: Conceptualization, H.Z. and C.X.; methodology, H.Z. and C.X.; software, H.Z.; validation, C.X.; formal analysis, Y.F.; investigation, H.T.; resources, C.X.; data curation, C.X.; writing—original draft preparation, H.Z. and C.X.; writing—review and editing, H.Z., C.X., Y.F. and H.T.; visualization, C.X.; supervision, C.X.; project administration, C.X.; funding acquisition, C.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported in part by the National Natural Science Foundation of China under Grant 61901221 and Grant 52005265, and in part by the National Key Research and Development Program of China under Grant 2019YFD1100404.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Chen, Q.; Song, H.; Yu, J.; Kim, K. Current development and applications of super-resolution ultrasound imaging. *Sensors* **2021**, *21*, 2417.
2. Zhou, H.; Zhuang, Z.; Liu, Y.; Liu, Y.; Zhang, X. Defect classification of green plums based on deep learning. *Sensors* **2020**, *20*, 6993.
3. Yan, X.; Liu, Y.; Xu, Y.; Jia, M. Multistep forecasting for diurnal wind speed based on hybrid deep learning model with improved singular spectrum decomposition. *Energ. Convers. Manage.* **2020**, *225*, 113456.
4. Pan, Z.; Tan, Z.; Lv, Q. A deep multi-frame super-resolution network for dynamic scenes. *Appl. Sci.* **2021**, *11*, 3285.
5. Xie, C.; Zeng, W.; Lu, X. Fast single-image super-resolution via deep network with component learning. *IEEE T Circ. Syst. Vid.* **2019**, *29*, 3473–3486.
6. Gu, J.; Wang, Z.; Kuen, J.; Ma, L.; Shahroudy, A.; Shuai, B.; Liu, T.; Wang, X.; Wang, G.; Cai, J.; et al. Recent advances in convolutional neural networks. *Pattern Recogn.* **2018**, *77*, 354–377, doi:10.1016/j.patcog.2017.10.013.
7. Minaee, S.; Boykov, Y.Y.; Porikli, F.; Plaza, A.J.; Kehtarnavaz, N.; Terzopoulos, D. Image segmentation using deep learning: A survey. *IEEE T Pattern Anal.* **2021**, doi:10.1109/TPAMI.2021.3059968.
8. Garcia-Garcia, A.; Orts-Escolano, S.; Oprea, S.; Villena-Martinez, V.; Martinez-Gonzalez, P.; Garcia-Rodriguez, J. A survey on deep learning techniques for image and video semantic segmentation. *Appl. Soft Comput.* **2018**, *70*, 41–65, doi:10.1016/j.asoc.2018.05.018.
9. Bouwmans, T.; Javed, S.; Sultana, M.; Jung, S.K. Deep neural network concepts for background subtraction:A systematic review and comparative evaluation. *Neural Netw.* **2019**, *117*, 8–66.
10. Yao, G.; Lei, T.; Zhong, J. A review of convolutional-neural-network-based action recognition. *Pattern Recogn. Lett.* **2019**, *118*, 14–22.
11. Di Wu; Zheng, S.; Zhang, X.; Yuan, C.; Cheng, F.; Zhao, Y.; Lin, Y.; Zhao, Z.; Jiang, Y.; Huang, D. Deep learning-based methods for person re-identification: A comprehensive review. *Neurocomputing* **2019**, *337*, 354–371.
12. Liu, W.; Wang, Z.; Liu, X.; Zengb, N.; Liu, Y.; Alsaadi, F. A survey of deep neural network architectures and their applications. *Neurocomputing* **2017**, *234*, 11–26, doi:10.1016/j.neucom.2016.12.038.
13. Zheng, S.; Song, Y.; Leung, T.; Goodfellow, I. Improving the robustness of deep neural networks via stability training. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp 4480–4488.
14. Nouiehed, M.; Razaviyayn, M. Learning deep models: Critical points and local openness. *arXiv* **2018**, arXiv: 1803.02968.
15. Vidal, R.; Bruna, J.; Giryes, R.; Soatto, S. Mathematics of deep learning. *arXiv* **2017**, arXiv:1712.04741.
16. Giryes, R.; Sapiro, G.; Bronstein, A.M. On the stability of deep networks. *arXiv* **2014**, arXiv:1412.5896.
17. Yan, X.; Liu, Y.; Huang, D.; Jia, M. A new approach to health condition identification of rolling bearing using hierarchical dispersion entropy and improved laplacian score. *Struct. Health Monit.* **2020**, doi:10.1177/1475921720948620.
18. Yan, X.; Liu, Y.; Jia, M. Health condition identification for rolling bearing using a multi-domain indicator-based optimized stacked denoising autoencoder. *Struct Health Monit.* **2020**, *19*, 1602–1626.
19. Huang, Y.; Si, W.; Chen, K.; Sun, Y. Assessment of tomato maturity in different layers by spatially resolved spectroscopy. *Sensors* **2020**, *20*, 7229.
20. Lei, W.; Jiang, X.; Xu, L.; Luo, J.; Xu, M.; Hou, F. Continuous Gesture Recognition Based on Time Sequence Fusion Using MIMO Radar Sensor and Deep Learning. *Electronics* **2020**, *9*, 869, doi:10.3390/electronics9050869.
21. Muhammad, W.; Aramvith, S. Multi-Scale Inception Based Super-Resolution Using Deep Learning Approach. *Electronics* **2019**, *8*, 892, doi:10.3390/electronics8080892.

22. Sun, Y.; Shi, Y.; Yang, Y.; Zhou, W. Perceptual Metric Guided Deep Attention Network for Single Image Super-Resolution. *Electronics* **2020**, *9*, 1145, doi:10.3390/electronics9071145.
23. Xie, C.; Zeng, W.L.; Jiang, S.Q.; Lu, X.B. Multiscale self-similarity and sparse representation based single image super-resolution. Neurocomputing 2017, 260, 92-103, doi:10.1016/j.neucom.2017.03.073.
24. Xie, C.; Liu, Y.; Zeng, W.; Lu, X. An improved method for single image super-resolution based on deep learning. *Signal Image Video Process.* **2019**, *13*, 557–565.
25. Yan, X.; Liu, Y.; Xu, Y.; Jia, M. Multichannel fault diagnosis of wind turbine driving system using multivariate singular spectrum decomposition and improved Kolmogorov complexity. *Renew. Energ.* **2021**, *170*, 724–748.
26. Du, J.; Cheng, K.; Yu, Y.; Wang, D.; Zhou, H. Panchromatic Image super-resolution via self attention-augmented wasserstein generative adversarial network. *Sensors* **2021**, *21*, 2158.
27. Alam, M.S.; Kwon, K.; Erdenebat, M.; Abbass, M.Y.; Alam, M.A.; Kim, N. Super-resolution enhancement method based on generative adversarial network for integral imaging microscopy. *Sensors* **2021**, *21*, 2164.
28. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
29. Bae, A.; Kim, W. Speaker Verification Employing Combinations of Self-Attention Mechanisms. *Electronics* **2020**, *9*, 2201, doi:10.3390/electronics9122201.
30. Zhang, Y.; Li, K.; Li, K.; Wang, L.; Zhong, B.; Fu, Y. Image super-resolution using very deep residual channel attention networks. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 286–301.
31. Park, S.C.; Park, M.K.; Kang, M.G. Super-resolution image reconstruction: A technical overview. *IEEE Signal Proc. Mag.* **2003**, *20*, 21–36.
32. Ha, V.K.; Ren, J.C.; Xu, X.Y.; Zhao, S.; Xie, G.; Masero, V.; Hussain, A. Deep Learning Based Single Image Super-resolution: A Survey. *Int. J. Autom. Comput.* **2019**, *16*, 413–426.
33. Anwar, S.; Khan, S.; Barnes, N. A deep journey into super-resolution: A survey. *ACM Comput. Surv.* **2020**, *53*, 1–34.
34. Yang, Z.; Shi, P.; Pan, D. A survey of super-resolution based on deep learning. In Proceedings of the 2020 International Conference on Culture-Oriented Science & Technology (ICCST), Beijing, China. 30–31 October 2020.
35. Wang, Z.; Chen, J.; Hoi, S.C.H. Deep learning for image super-resolution: A survey. *IEEE T Pattern Anal.* **2020**, doi: 10.1109/TPAMI.2020.2982166.
36. Kim, S.; Jun, D.; Kim, B.; Lee, H.; Rhee, E. Single image super-resolution method using cnn-based lightweight neural networks. *Appl. Sci.* **2021**, *11*, 1092.
37. Liu, Y.; Zhang, G.; Wang, H.; Zhao, W.; Zhang, M.; Qin, H. An Efficient Super-Resolution Network Based on Aggregated Residual Transformations. *Electronics* **2019**, *8*, 339, doi:10.3390/electronics8030339.
38. Du, J.; Han, M.; Jin, L.; Hua, Y.; Li, S. Target Localization Methods Based on Iterative Super-Resolution for Bistatic MIMO Radar. *Electronics* **2020**, *9*, 341, doi:10.3390/electronics9020341.
39. Shi, Y.; Li, B.; Wang, B.; Qi, Z.; Liu, J. Unsupervised Single-Image Super-Resolution with Multi-Gram Loss. *Electronics* **2019**, *8*, 833, doi:10.3390/electronics8080833.
40. Sahito, F.; Zhiwen, P.; Ahmed, J.; Memon, R.A. Wavelet-Integrated Deep Networks for Single Image Super-Resolution. *Electronics* **2019**, *8*, 553, doi:10.3390/electronics8050553.
41. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
42. Kim, J.; Lee, J.K.; Lee, K.M. Accurate image super-resolution using very deep convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1646–1654.
43. Lim, B.; Son, S.; Kim, H.; Nah, S.; Mu Lee, K. Enhanced deep residual networks for single image super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Honolulu, HI, USA, 21–26 July 2017; pp. 136–144.
44. Wang, J.; Teng, G.; An, P. Video Super-Resolution Based on Generative Adversarial Network and Edge Enhancement. *Electronics* **2021**, *10*, 459, doi:10.3390/electronics10040459.
45. Ooi, Y.K.; Ibrahim, H. Deep Learning Algorithms for Single Image Super-Resolution: A Systematic Review. *Electronics* **2021**, *10*, 867.
46. Xie, C.; Zeng, W.L.; Jiang, S.Q.; Lu, X.B. Bidirectionally aligned sparse representation for single image super-resolution. Multimed. Tools Appl. 2018, 77, 7883-7907, doi:10.1007/s11042-017-4689-7.
47. Tao, D.; Jianrui, C.; Zhang, Y.B.; Xia, S.-T. Second-order attention network for single image super-resolution. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 11057–11066.
48. Anwar, S.; Barnes, N. Densely residual laplacian super-resolution. *IEEE T Pattern Anal.* **2020**, doi:10.1109/TPAMI.2020.3021088.
49. Yang, C.; Lu, G. Deeply recursive low- and high-frequency fusing networks for single image super-resolution. *Sensors* **2020**, *20*, 7268.
50. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
51. Agustsson, E.; Timofte, R. Ntire 2017 challenge on single image super-resolution: Dataset and study. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Honolulu, HI, USA, 21–26 July 2017; pp. 126–135.

52. Zhang, K.; Zuo, W.; Gu, S.; Zhang, L. Learning deep CNN denoiser prior for image restoration. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA, 21–26 July 2017; pp. 3929–3938.
53. Li, P.; Xie, J.; Wang, Q.; Zuo, W. Is second-order information helpful for large-scale visual recognition? In Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy, 22–29 October 2017; pp. 2070–2078.
54. Benesty, J. Adaptive eigenvalue decomposition algorithm for passive acoustic source localization. *J. Acoust. Soc. Am.* **2000**, *107*, 384–391.
55. Higham, N.J. Functions of matrices: Theory and computation. In Proceedings of the SIAM, Atlanta, GA, USA 24–26 April 2008.
56. Musunuri, Y.R.; Kwon, O.-S. Deep Residual Dense Network for Single Image Super-Resolution. *Electronics* **2021**, *10*, 555, doi:10.3390/electronics10050555.
57. Choi, J.; Kim, M. A deep convolutional neural network with selection units for super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, USA, 21–26 July 2017; pp. 1150–1156.
58. Liu, J.; Zhang, W.; Tang, Y.; Tang, J.; Wu, G. Residual feature aggregation network for image super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, WA, USA, 14–19 June, 2020; pp. 2359–2368.
59. Hu, Y.; Li, J.; Huang, Y.; Gao, X. Channel-wise and spatial feature modulation network for single image super-resolution. *IEEE T Circ. Syst. Vid.* **2019**, *30*, 3911–3927.
60. Kim, J.; Choi, J.; Cheon, M.; Lee, J. RAM: Residual attention module for single image super-resolution. *arXiv Comput. Vis. Pattern Recognit.* **2018**, arXiv:1811.12043.
61. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7794-7803.
62. Zhang, Y.; Li, K.; Li, K.; Zhong, B.; Fu, Y. Residual non-local attention networks for image restoration. *arXiv* **2019**, arXiv:1903.10082 2019.
63. Mei, Y.; Fan, Y.; Zhou, Y.; Huang, L.; Huang, T.S.; Shi, H. Image super-resolution with cross-scale non-local attention and exhaustive self-exemplars mining. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 5690–5699.
64. Mei, Y.; Fan, Y.; Zhang, Y.; Yu, J.; Zhou, Y.; Liu, D.; Fu, Y.; Huang, T.S.; Shi, H. Pyramid attention networks for image restoration. *arXiv* **2020**, arXiv:2004.13824.
65. Hou, Q.; Zhou, D.; Feng, J. Coordinate Attention for efficient mobile network design. *arXiv* **2021**, arXiv:2103.02907.
66. Huang, Y.; Lu, R.; Chen, K. Detection of internal defect of apples by a multichannel Vis/NIR spectroscopic system. *Postharvest Biol. Tec.* **2020**, *161*, 111065.
67. Yan, X.; Liu, Y.; Jia, M. Research on an enhanced scale morphological-hat product filtering in incipient fault detection of rolling element bearings. *Measurement* **2019**, *147*, 106856.
68. Bevilacqua, M.; Roumy, A.; Guillemot, C.; Alberi-Morel, M.L. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In Proceedings of the BMVC, Surrey, UK, 3–7 September 2012.