applied sciences

MDPI

Article

Learnable Nonlocal Contrastive Network for Single Image Super-Resolution

Binbin Xu * and Yuhui Zheng

Engineering Research Center of Digital Forensics, Ministry of Education, Nanjing University of Information Science and Technology, Nanjing 210044, China; zheng_yuhui@nuist.edu.cn
* Correspondence: 20211249460@nuist.edu.cn

Abstract: Single image super-resolution (SISR) aims to recover a high-resolution image from a single low-resolution image. In recent years, SISR methods based on deep convolutional neural networks have achieved remarkable success, and some methods further improve the performance of the SISR model by introducing nonlocal attention into the model. However, most SISR methods that introduce nonlocal attention focus on more complex attention mechanisms and only use fixed functions for measurement when exploring image similarity. In addition, the model penalizes the algorithm in terms of loss when the output predicted by the model does not match the target data, even if this output is a potentially valid solution. To this end, we propose learnable nonlocal contrastive attention (LNLCA), which flexibly aggregates image features while maintaining linear computational complexity. Then, we introduce the adaptive target generator (ATG) model to address the problem of the single model training mode. Based on LNLCA, we construct a learnable nonlocal contrastive network (LNLCN). The experimental results demonstrate the effectiveness of the algorithm, which produces reconstructed images with more natural texture details.

Keywords: single image super-resolution; deep convolutional network; nonlocal attention mechanism; adaptive target

Citation: Xu, B.; Zheng, Y. Learnable Nonlocal Contrastive Network for Single Image Super-Resolution. *Appl. Sci.* **2023**, *13*, 7160. https://doi.org/10.3390/app13127160

Academic Editor: Yudong Zhang

Received: 15 May 2023
Revised: 13 June 2023
Accepted: 13 June 2023
Published: 15 June 2023

1. Introduction

As a classic image restoration problem, single image super-resolution (SISR) aims to restore a given low-resolution (LR) image into a detailed high-resolution (HR) image. Because of its extensive application value in remote-sensing imaging [1,2], video surveillance [3,4], pedestrian detection [5], and other fields [6,7], it has attracted the attention of many researchers. However, due to the irreversibility of the image degradation process, multiple potential HR image solutions may be derived from a single input LR image. Therefore, image super-resolution reconstruction is essentially an ill-posed problem.

To solve this ill-posed problem, researchers have proposed many algorithms. Traditional SISR algorithms are roughly divided into three categories, namely, interpolation-based SISR [8–10], reconstruction-based SISR [11], and shallow-based SISR [12]. With the development of convolutional neural networks (CNN), CNN-based SISR methods have been widely used in the field of image super-resolution and achieved ideal results. Methods of this type can flexibly learn the deep features of images and establish a mapping relationship between an input image and an output image, and they have achieved significant improvements compared to the traditional method. SRCNN [13] applies CNNs to the SISR task for the first time, achieving impressive performance. After that, many research scholars proposed SISR methods with deeper CNN layers. Considering that the low-frequency information contained in LR images and HR images has great similarity, some scholars have significantly improved the performance of the network models by introducing the idea of residual structure. Kim et al. [14] alleviated the slow training problem of deep CNN by introducing residual learning and proposed a more efficient deep CNN-based model,

VDSR. Although CNN-based SISR methods have made significant progress, most of them treat different image layers between networks equally, ignore the potential image feature correlation between adjacent network layers, which reduces the representation ability of deep network models.

The attention mechanism can allocate resources to the input area with a large amount of information. Recently, scholars have greatly improved the performance of CNN models by applying attention mechanisms to super-resolution tasks [15–19]. Zhang et al. [15] proposed the residual channel attention network (RCAN), which has improved discriminative learning ability for cross-channel features due to the introduction of channel attention and produces higher-quality SR images. Since nonlocal attention can utilize the intrinsic feature correlations in images, scholars have introduced a nonlocal attention into the deep networks to further improve the performance. The SAN network model proposed by Dai et al. [20] exploits image high-order channel information and image self-similarity to discover the intrinsic correlations between different network layers, thereby capturing the potential intrinsic information of the image. Reviewing these classic methods, we find that most attention-based SR methods are dedicated to studying more complex attention mechanisms, which makes their algorithms increasingly demanding for training. Some methods apply nonlocal attention to reduce the huge computational overhead imposed by modules by limiting the search range of the modules, which causes the model to ignore important globally relevant information and be insufficiently flexible. Furthermore, most models use only the HR images of a given training set as the mappings for the LR image training, which may ignore potentially better solutions and limit the generalization of the model.

To address the above issues, we proposed a learnable nonlocal contrastive network (LNLCN), which can explore the correlations between image intrinsic features and different channel features of the network and obtain SR results with clearer textures. Inspired by efficient nonlocal contrastive attention (ENLCA) [21] and global learnable attention (GLA) [22], we proposed learnable nonlocal contrastive attention (LNLCA), which obtains linear computational complexity through the combination of similarity functions and matrix multiplication and can more effectively aggregate relevant information within an image. To improve the ill-posedness of SR tasks, we introduced an adaptive target generator (ATG) [23] module to relax the model's constraints on potential solutions, thereby generating more natural SR results. In addition, inspired by CutBlur [24], we proposed a new data augmentation method (RFP) for the training data that achieves data augmentation without destroying image pixel correlation. Our LNLCN achieves sharper visual SR results than state-of-the-art SISR methods.

Overall, the main contributions of our work are listed as follows:

1. A novel learnable nonlocal contrastive attention (LNLCA) method was proposed for exploring nonlocal textures with low similarity but more precise details, which maintains linear computational complexity while aggregating important image features;
2. We proposed a deep feature fusion attention group (DFFAG) for fusing local adjacency information and learnable nonlocal self-similar information, thereby helping the network repair damaged texture regions;
3. We introduced the adaptive target generator (ATG), which can alleviate the ill-posedness of the SR task and further explore potential solutions by endowing the model with more output flexibility.

The remainder portions of the paper are structured as follows: Section 2 presents recent related work in the field of single-image super-resolution. Section 3 presents the proposed algorithm in detail. Section 4 evaluates the performance of the proposed model through tests on standard datasets. Section 5 discusses the complexity and computation of the model. Section 6 concludes the paper and discusses future research.

2. Related Work

2.1. CNN-Based Methods

Since the convolutional neural networks can effectively represent the nonlinear mapping between LR images and HR images, they have been widely used to solve the SISR problem. Dong et al. [13] pioneered the application of a CNN to SISR, and the proposed SRCNN achieved impressive performance with a three-layer convolutional layer as the network architecture. Shi et al. [25] proposed an ESPCN that can reconstruct images in real time, which extracts features from the input LR images and further applies subpixel convolutional layers embedded to upsample the reconstructed images. Later, this method of upsampling with subpixel convolutional layers at the end of the network architecture became the primary choice for subsequent SISR network architectures. Some methods improve the model's ability to represent image features by increasing the network size. Kim et al. alleviated the problems of model training difficulty and slow convergence, by introducing residual strategies, and proposed VDSR [14] and DRCN [26] with deep network architectures. Tong et al. [27] proposed an efficient network, namely, SRDenseNet, by introducing dense blocks to maintain connections between different network layers. The EDSR proposed by Lim et al. [28] achieves better performance metrics than previous network models by improving the residual structure. Perceptual quality is an important evaluation metric for reconstructed images, Ledig et al. [29] proposed the SRGAN, which significantly improves the texture realism of SR images by using generative adversarial networks and utilizing a multitask loss. When compared to conventional classic algorithms, the majority of these methods have shown significant improvement, but, since they ignore the feature correlations of intermediate network layers, it is difficult to reconstruct SR results with natural texture details.

2.2. Attention-Based Methods

An attention mechanism enables an SR network focus on important information, thereby helping the network to distinguish relevant information that is beneficial for reconstruction. Recently, scholars have greatly improved the evaluation indicators of deep network models by introducing attention mechanisms. SENet, which was proposed by Hu et al. [30], has greatly improved the accuracy of image classification models by introducing network channel features. Inspired by this, RCAN, proposed by Zhang et al. [15], combines a residual block with a channel attention. The network model is more than 400 layers deep and has achieved significant improvements in image reconstruction quality. SAN, proposed by Dai et al. [20], uses second-order channel attention (SOCA) for correlation learning while also employing the nonlocal enhancement residual group (NLRG) to capture the self-similarity information of the input image. To reduce the huge amount of calculations required by the nonlocal attention mechanism, scholars have investigated efficient nonlocal attention mechanisms. Mei et al. [31] proposed an NLSN network model that uses locality-sensitive hashing (LSH) to identify the most important information regions, which improves the computational efficiency and performance metrics of the model for nonlocal attention. Xia et al. [21] adopted the kernel method to approximate the computational load of the nonlocal module method as an exponential function, introduced contrastive learning to make the model focus on the sparse aggregation of image information, and proposed an efficient contrastive attention model (ENLCN). Su et al. [22] proposed a deep-learnable similarity network (DLSN) that explores the self-similarity in nonlocal textures, mines nonlocal textures with low similarity but more precise details, and restores damaged textures. Lu et al. [32] proposed a super-resolution network model (ESRT) which realizes long-term image information modeling by combining CNN and self-attention in transformer and uses feature segmentation and high-frequency filter modules (HFM) to improve model calculation efficiency. Chen et al. [33] proposed a deep multi-stage network (MAAN) for accurate image SR by stacking attention enhancement modules and making full use of the advantages of different modules. Fang et al. [34] proposed a hybrid network of CNN

and Transformer (HNCT) for lightweight SISR, which extracts deep features that are more conducive to image SR by exploiting local and nonlocal priors.

2.3. Multiple-Chioce Learning

Ensemble learning aims to reduce the error correlation by combining the diversity among training member models under the joint loss, thereby improving the performance of the model. The research on ensemble learning by Salamo et al. [35] and Krogh et al. [36] demonstrated that cross-member model training can improve performance, which laid the theoretical foundation for neural network model ensemble learning. Later, the work of Guzman-Rivera et al. [37–39] on multiple-choice learning (MCL) provided another attractive scheme that encourages the loss to generate diverse outputs by comparing the predicted output of a single solution with the outputs of different member models. Jo et al. [23] proposed an ATG training strategy suitable for SISR tasks. Unlike the MCL method, ATG uses multiple potential solutions to generate a prediction to further train the model as a loss.

3. Method

This section introduces our learnable nonlocal contrastive network (LNLCN). The structure of LNLCN is shown in Figure 1. Next, we present the learnable nonlocally residual group (LNLRG), adaptive target generator (ATG), and RFP data augmentation methods in detail.

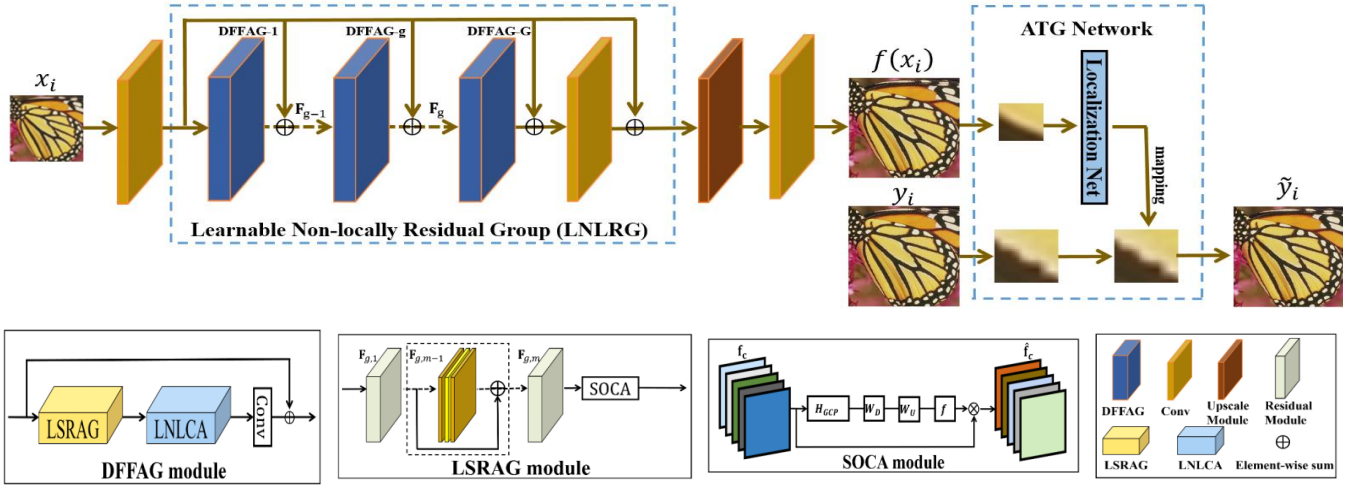

Figure 1. Architecture of the proposed learnable nonlocal contrastive network (LNLCN).

3.1. Network Structure

Our proposed LNLCN consisted of five main modules: a shallow feature extraction module, a deep feature extraction module based on the learnable nonlocally residual group (LNLRG), an upsampling module, a reconstruction module, and an adaptive target generator module. We represented by I_{LR} and I_{SR} the input image and output reconstructed image of LNLCN, respectively. First, we extracted feature information on the input I_{LR} through a convolutional layer:

$$F_0 = H_{SF}(I_{LR}) \tag{1}$$

where $H_{SF}(*)$ represents the operation of a convolution layer. Then, the obtained image features F_0 were input into the learnable nonlocally residual group (LNLRG) to extract image depth features:

$$F_{DF} = H_{LNLRG}(F_0) \tag{2}$$

where $H_{LNLRG}(*)$ represents our proposed learnable nonlocally residual group (LNLRG), which consists of G deep feature fusion attention group (DFFAG) models that focus on

image-related information. After that, we upscaled the resulting deep feature F_{DF} through the upsampling module:

$$F_{UP} = H_{UP}(F_{DF}) \tag{3}$$

where $H_{UP}(*)$ and F_{UP} represent the upsampling module and the upsampled features, respectively. Some effective upsampling modules, such as transposed convolution [40] and subpixel convolution [25], have been shown to improve performance without increasing the computational complexity and, thus, are more suitable for deep network models. After that, we reconstructed F_{UP} through a convolutional layer to obtain an SR image:

$$I_{SR} = H_{SR}(F_{UP}) = H_{LNLCN}(I_{LR}) \tag{4}$$

where $H_{SR}(*)$ and $H_{LNLCN}(*)$ represent the features of the reconstruction module and the proposed LNLCN, respectively.

After that, we optimized the model with a loss function. We used the same L1 loss as in previous work (SAN) [20]. Specifically, for a given training set $\left\{I_{LR}^{i}, I_{HR}^{i}\right\}_{i=1}^{N}$ containing N pairs of LR images I_{LR} to be processed and corresponding original HR images I_{HR}, the ultimate goal of model training was to minimize the L1 loss function:

$$L(\Theta) = \frac{1}{N}\sum_{i=1}^{N} \parallel H_{LNLCN}\left(I_{LR}^{i}\right) - I_{HR}^{i} \parallel_{1} \tag{5}$$

where Θ represents the parameter set of the LNLCN. In the optimization of the loss function, we used the stochastic gradient descent algorithm to ensure the stability of the training. After the initial training of the LNLCA network, an adaptive target $\widetilde{y}_i$ was generated through the ATG network:

$$\widetilde{y}_i = ATG(I_{HR}, H_{LNLCN}(I_{LR})) \tag{6}$$

where $ATG(*)$ and $H_{LNLCN}(I_{LR})$ represent the adaptive target generation network and the current prediction output of the LNLCN, respectively. The ATG network affinely transformed I_{HR} to $H_{LHAN}(I_{LR})$ within an acceptable range. Finally, the pretrained model was further fine-tuned by the adaptive objective $\widetilde{y}_i$ generated from the ATG network:

$$\sum_{i} \ell\left(\widetilde{y}_i, H_{LHAN}(I_{LR})\right) \tag{7}$$

where $\ell(*)$ represents the L2 loss. The ATG network could adaptively generate additional target $\widetilde{y}_i$ according to the current iterative prediction of the pretrained network; hence, we did not have to reprocess the training data, and it did not take a long time.

3.2. Learnable Nonlocally Residual Group (LNLRG)

We now present the learnable nonlocally residual group (LNLRG) models in detail, which consisted of G deep feature fusion attention group (DFFAG) models. Specifically, each DFFAG module consisted of a local source residual attention group (LSRAG) [20] and a learnable nonlocal contrastive attention module (LNLCA). The DFFAG module can integrate local and global features, and the shared connection architecture design can better extract and fuse high-frequency features, making the model more accurate and efficient when processing complex data, thereby improving the performance of the model.

3.2.1. Deep Feature Fusion Attention Group

To achieve efficient and accurate image feature extraction and fusion, we combined a local source residual attention group (LSRAG) [20] with an innovative learnable nonlocal contrastive attention module (LNLCA) to construct a deep feature fusion attention group (DFFAG) module. Different DFFAG modules use the shared source residual skipping (SSC)

connection [20], which ensures that the model bypasses the low-frequency information of the training samples to the greatest extent possible, thereby focusing on the training of high-frequency features. This design ensures that the LNLRG model has higher efficiency and accuracy in the process of image feature extraction. Among them, the g-th DFFAG module can be expressed as follows:

$$F_g = W_{SSC}F_0 + H_g(F_{g-1}) \tag{8}$$

where W_{SSC} represents the parameters of the convolutional kernel, which is initially set to 0 and then updated throughout the iteration of the network layer, and $H_g(*)$, F_{g-1}, and F_g represent the function of the g-th DFFAG module and the input features and output features of the DFFAG module, respectively. Furthermore, the depth features F_{DF} of the image can be expressed as below:

$$F_{DF} = W_{SSC}F_0 + F_G \tag{9}$$

This method of stacking and simplifying residual blocks helps form a deep CNN and accelerates the training of a network model with high-performance reconstructed images.

As previously discussed in the work on SAN [20], we introduced the LSRAG module to capture the channel features of images and fully mine the feature correlations between image channels (see Figure 2). Taking the g-th LSRAG module as an example, the function of the m-th residual block in LSRAG module can be expressed as shown:

$$F_{g,m} = H_{g,m}(F_{g,m-1}) \tag{10}$$

where $F_{g,m-1}$, $F_{g,m}$ represent the inputs and outputs of the LSRAG module. We input the obtained $F_{g,m}$ into the SOCA [20] module to obtain channel attention. The output of the DFFAG module was further obtained, which can be expressed as follows:

$$F_g = H_{conv}(H_{LNLCA}(H_{SOCA}(F_{g,m}))) \tag{11}$$

where H_{SOCA} and H_{LNLCA} represent the functions of the SOCA module in the g-th LSRAG and LNLCA modules, respectively. Further, the output F_g of the g-th DFFAG module can be obtained through a convolutional layer. Our proposed LNLCA module can fully exploit more valuable nonlocal textures in LR images by modifying the self-similarity function, enhancing the long-range modelling capability of the model.

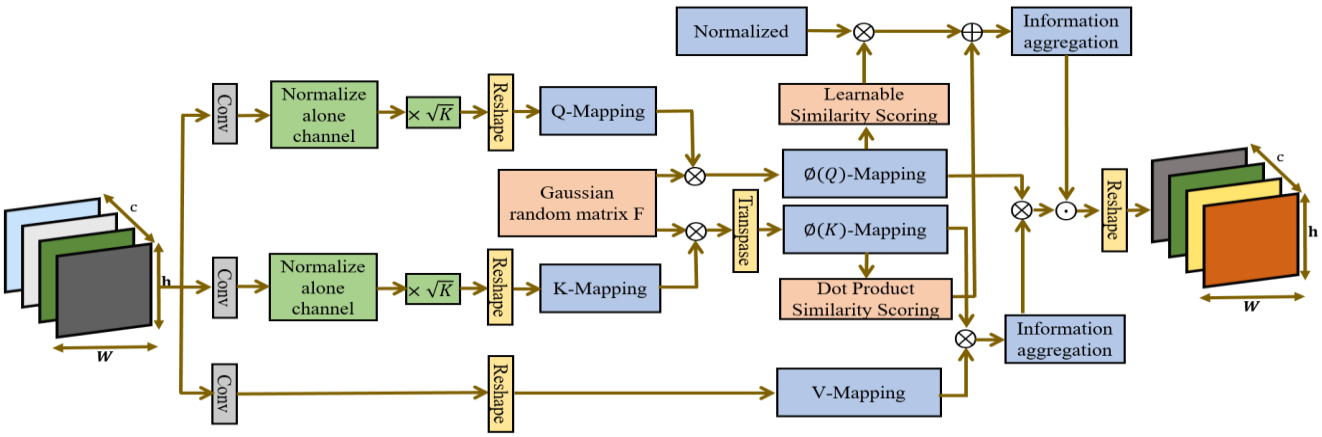

Figure 2. Structure of LNLCA. h, w, and c represent the training image height, width, and dimension, respectively. m and N represent the number of random samples and the size of the input feature map, respectively. Q, K, and V represent three sets of mappings. Among them, Q and K are further transformed into $\phi(Q)$ and $\phi(K)$ by Gaussian random matrix F.

3.2.2. Learnable Nonlocal Contrastive Attention

As shown in Figure 2, the proposed LNLCA module combines the advantages of self-similar exploration and sparse aggregation strategies and can achieve global correlation information aggregation of important information at the cost of linear complexity, avoiding overfitting while improving the computational efficiency of the model combination and information redundancy problem. Specifically, the self-similar exploration strategy can find feature information highly related to the current location in the local area and aggregate it with the current location. The sparse aggregation strategy can effectively reduce the computational complexity while maintaining the integrity of global information.

Formally, the process by which the LNLCA module applies attention to feature vectors is defined as shown:

$$y_i = \sum_{j=1}^{n} \frac{exp\left(s\left(Q_i, K_j\right)\right)}{\sum_{\hat{j}=1}^{n} f\left(s\left(Q_i, K_{\hat{j}}\right)\right)} v_j \tag{12}$$

where Q_i, K_j, and $K_{\hat{j}}$ represent the pixel features corresponding to positions i, j, and $\hat{j}$, respectively, of feature map X; y_i represents the output for position i; $f(*)$ represents the similarity measurement between feature map Q and feature map K; v_j represents the feature transformation function; and n represents the input size. Among them, to explore more valuable nonlocal textures, we proposed to incorporate the learnable similarity score function (LSS) [22] and the dot product similarity score function (DPSS) [22] into the LNLCA module, thereby forming the $s(*,*)$ function. Furthermore, different from traditional nonlocal attention mechanisms, our LNLCA module achieves better performance by approximating Gaussian random features and changing the multiplication order. Formally, the $s(*,*)$ function can be defined as below:

$$s\left(Q_i, K_j\right) = s_l^j(Q_i) + s_f\left(Q_i, K_j\right) \tag{13}$$

where $s_l^j(*)$ represents the score of position j in the learnable similarity scoring function (LSS) and $s_f(*)$ represents the dot product similarity scoring function (DPSS). $s_l(*)$ can adaptively modify the scores of different positions in the nonlocal texture, thereby helping the network correct some inaccurate textures, which is formally defined as follows:

$$s_l(Q_i) = W_2\sigma(W_1\theta(Q_i) + b_1) + b_2 \tag{14}$$

where $\sigma(*)$ represents the ReLU function and $W_1 \in X^{n\times c}$, $W_2 \in X^{n\times n}$, $b_1 \in X^n$, and $b_2 \in X^n$. The dot product similarity score $s_f(*)$ is formally defined as shown:

$$s_f\left(Q_i, K_j\right) = \theta(Q_i)^T\delta\left(K_j\right) \tag{15}$$

For an input feature map X, we first multiplied by a scaling factor k to increase the weight of the relevant information:

$$Q = \sqrt{k}\frac{\theta(X)}{\| \theta(X) \|}, K = \sqrt{k}\frac{\delta(X)}{\| \theta(X) \|}, V = g(X) \tag{16}$$

$$\mathrm{K}\left(Q_i, K_j\right) = \exp\left(Q_i^\top K_j\right) = \phi(Q_i)^T\phi\left(K_j\right) \tag{17}$$

where X and k represent the module input feature map and the magnification factor, respectively; $\theta(*)$, $\delta(*)$, and $g(*)$ represent feature transformation functions; and Q_i and K_j represent the pixel information at positions i and j in the feature maps Q and K, respectively. By setting r Gaussian random samples and concatenating these samples into a Gaussian random matrix F, the Gaussian random feature map was further approximated

as $\phi(Q_i)^T\phi(K_j)$, where $\phi(u) = \frac{1}{\sqrt{r}}\exp(-\parallel u \parallel^2 /2)\exp(Fu)$. Finally, an efficient nonlocal attention mechanism was obtained:

$$\hat{Y} = D^{-1}\left(\phi\left(Q\right)^\top\left(\phi(K)V^\top\right)\right) \tag{18}$$

$$D = diag\left[\phi\left(Q\right)^\top(\phi(K)1_N)\right] \tag{19}$$

where $\hat{Y}$ represents the obtained approximate nonlocal attention and D represents the normalization term in the softmax function.

3.3. Adaptive Target Generator (ATG)

Most SR models are trained by seeking a mapping between LR samples and HR samples in the training dataset. When the image predicted by the model does not fit the ground truth (GT) image, even if the output is a potential valid solution, the model is also penalized according to the training loss function. Inspired by the ATG network [23], we alleviated the ill-posedness of this SR task by introducing an adaptive target generator model to the SR model. By introducing the ATG network, the model was given some flexibility, thereby assisting the super-resolution model to generate more accurate reconstructed images. Different from the exploration of generative adversarial networks in [23], the proposed LNLCN model has extremely deep convolutional network layers, and we will explore the performance of ATG on deep convolutional architecture models. The adaptive object generation network (ATG) is presented in detail in Figure 3, which contains a localization network consisting of four fully connected layers. All layers, except the last one, contain the same number of BN layers and the ReLU function.

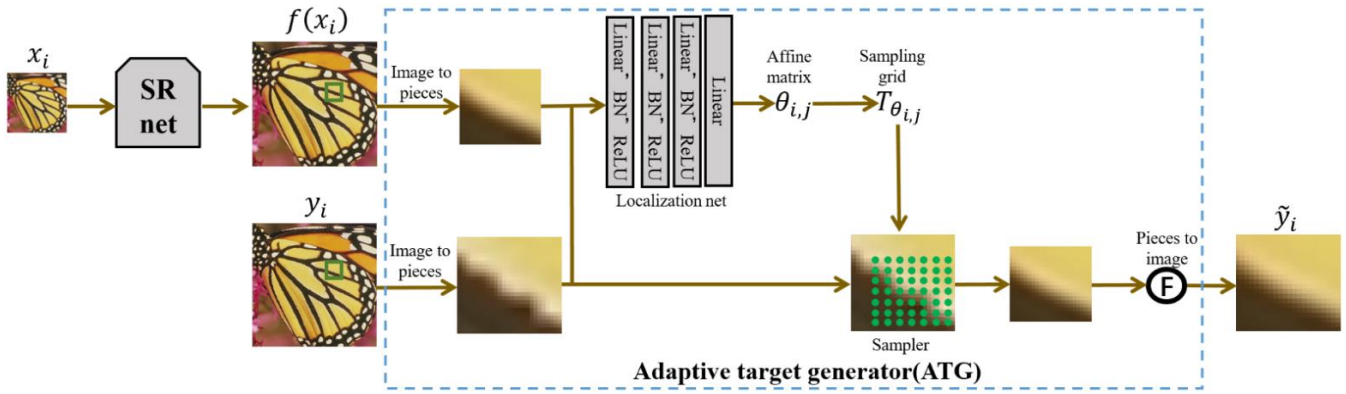

Figure 3. Structure of ATG. $f(x_i)$ and y_i represent the current prediction of the network and original real images. By converting each nonoverlapping part of y_i into a corresponding region of the network output $f(x_i)$, the adaptive target y is further rearranged to generate a new target.

During model training, the ATG network works in a patch-like fashion to adaptively generate new targets based on the predicted image for the current iteration of the SR model; therefore, there is no requirement for additional data preparation nor to conduct preprocessing. Specifically, the ATG network first slices the model prediction output $f(x_i)$ into nonoverlapping image slices with stride p and size p × p, and then slices the original ground truth (GT) image into overlapping image slices with stride p and size s × s, where p < s. The image patch size of GT is set slightly larger than the image patch size of $f(x_i)$ because this strategy needs to use the real target as the search space. Then, the affine transformation matrix $\theta_{i,j}$ is obtained through the positioning network, and the y_i image slices are further sent to the bilinear sampling network for feature change. The j-th y_i image slices are converted into corresponding $f(x_i)$ image slices, and, finally, our adaptive target is created by combining the transformed image patches of size p × p. Through the ATG network, we create a new target $\widetilde{y}_i$ that is most relevant to $f(x_i)$ predicted by

the SR network while keeping the original content unchanged, which is further used for network training.

3.4. RFP Data Augmentation

Data augmentation is an effective way to increase the performance metrics of a convolutional network. Most data augmentation methods increase prior knowledge by imposing constraints on the data, which can reduce the influence of negative information on the network and improve the accuracy of the model. However, SISR, as a class of classic low-level vision problems, has image pixels that are more sensitive to local and global relationships, and some image augmentation methods suitable for high-level vision tasks usually manipulate pixels or features, which hinders the reconstruction of images by super-resolution models. Traditional data augmentation strategies, such as rotation and flipping, have brought some benefits to SISR model performance. Inspired by the data expansion method CutBlur [24], we proposed an RFP data augmentation method. Specifically, the RFP data enhancement method performs image enhancement on each training image through a combined strategy of random rotation, horizontal flipping, and channel arrangement. Considering the sensitivity of the super-resolution problem to pixels, our RFP data enhancement method can arrange and mix the RGB channels in the original image (by adding constant values) without damaging the pixel space in the image. This enhancement method does not change the image structure and can provide good performance and synergy with other classic and traditional enhancement methods.

3.5. Implementation

For the network architecture, we followed the work of [20], adopted the LSRAG module of the SAN network as the cornerstone for construction, and combined the LNLCA module to form our deep feature fusion attention group (DFFAG). In the LNLCA module, we set the expansion factor to 6, edge b to 1, and the number of random samples r to 128. The learnable nonlocally residual group (LNLRG) consisted of G = 10 DFFAG modules, and we combined a second-order attention module (SOCA) with M = 10 residual blocks in each LSRAG module. In the ATG network, we set stride p = 7 and stride s = 9. For the upsampling module part, we adopted ESPCN [25] to upgrade the image depth features. Finally, a color SR image (RGB channel) was generated through a convolutional layer.

4. Experiment

4.1. Setup

After comparing RCAN [15], SAN [20], EDSR [28], and NLSN [31], we used DIV2K [41] as the training dataset for the network, which contains 800 samples of RGB images with a resolution of 2K. For model testing, we used five SISR benchmark datasets: Set5 [42], Set14 [43], B100 [44], Urban100 [45], and Manga109 [46]. All samples were analyzed with a bicubic downsampling (BI) degenerate model and a blurred descent (BD) model. We evaluated the predicted SR results using PSNR and SSIM, and all tests were performed on the Y channel after the image was converted to YCbCr space. During the model training process, we augmented the dataset images by RFP data augmentation. We provided 16 LR image patches with a size of 48 $\times$ 48 as the network input for each iteration batch. The LNLCA model was iteratively trained using the ADAM optimizer [47], where $\beta_1 = 0.9$, $\beta_2 = 0.99$, and $\epsilon = 10^{-8}$. During training, we initially set the learning rate to 10^{-4} and then halved it every 200 batches. We first trained the model for 8.4×10^5 iterations, after which we used the ATG training strategy to fine-tune the entire network for 5.6×10^4 iterations with a learning rate of 10^{-5}. In our LNLCN model, we used the ReLU activation function in most layers to help the model better learn nonlinear relationships between input data. At the same time, we also used some other activation functions, such as Sigmoid and Softmax. Our model was implemented in the PyTorch [48] on an Nvidia 1080Ti GPU.

4.2. Ablation Study of k in the LNLCA Module

In this study, the LNLCA module achieves superior performance with lower overhead than traditional nonlocal modules. To find a suitable value of the amplification factor k, we set different k values in the LNLCA module for the LNLCN model, and we conducted a more detailed ablation study. As shown in Figure 4, we explored the reasons for the effect of the scaling factor k on the existing infrastructure and evaluated the performance metrics of the model on five standard datasets.

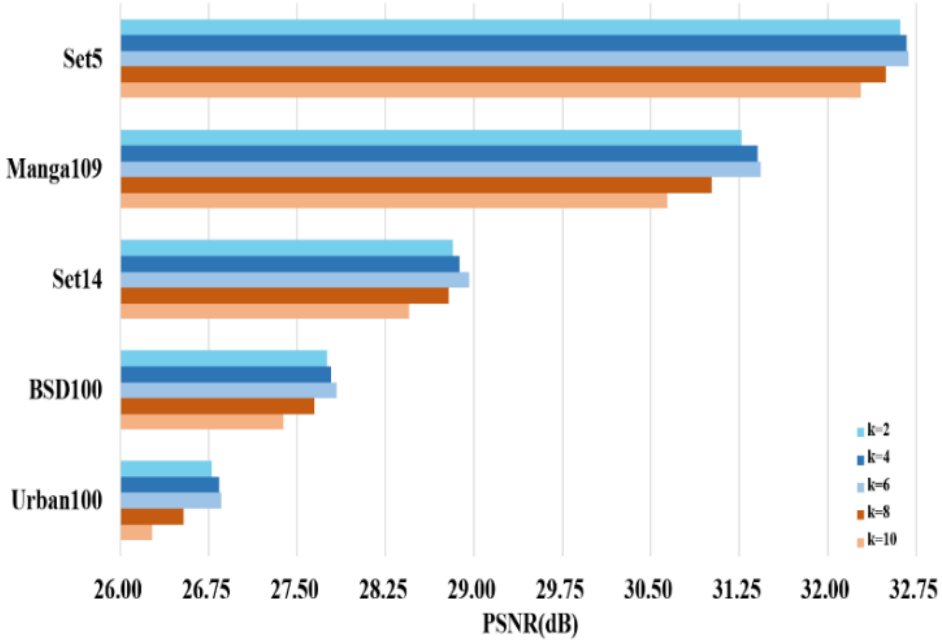

Figure 4. Ablation study using different values of k in LNLCA with the BI model (4×).

The PSNR value of the LNLCN model steadily increases with increasing k value. With a further increase in the amplification factor k, the performance of the model drops sharply. According to this experimental result, our inference is that increasing k causes the approximate variance of LNLCA to increase exponentially, which, in turn, leads to performance degradation. Therefore, setting k in the LNLCA module to six produces good results.

4.3. Ablation Study of the Number of ATG Iteration

To explore the impact of ATG network iteration number settings on the network performance, we conducted more detailed ablation experiments. Specifically, the number of iterations of the ATG network was set to 2.8×10^4, 4.2×10^4, 5.6×10^4, 7.0×10^4, 8.4×10^4, and 9.8×10^4. We further evaluated the performance of the model on the Set5 dataset with a downsampling factor of four. Table 1 shows the experimental results under different iterations. Through the data, we explored the best iteration settings for LNLCN. As shown in Figure 5, setting the number of iterations of the ATG network to 4.2×10^4 and 5.6×10^4 produces the best performance indicators, and the improvement in the SSIM values at this time is considerable. In addition, too many iterations, such as 9.8×10^4, lead to model performance degradation; this may be because too many iterations cause the model to overfat the dataset and affect the generalization ability of the model. Thus, in our LNLCN, we finally set the number of ATG network iterations to 5.6×10^4 to obtain the best reconstruction results. This result shows that our LNLCN model can effectively improve the quality and visual effect of super-resolution images with an appropriate number of iterations and has better practicality and application value.

Table 1. Iterative experiment of ATG network.

Iteration Settings	9.8×10^4	8.4×10^4	7.0×10^4	5.6×10^4	4.2×10^4	2.8×10^4
PSNR(dB)	32.267	32.435	32.559	32.638	32.625	32.608
SSIM	0.8943	0.8961	0.8982	0.9008	0.9002	0.8999

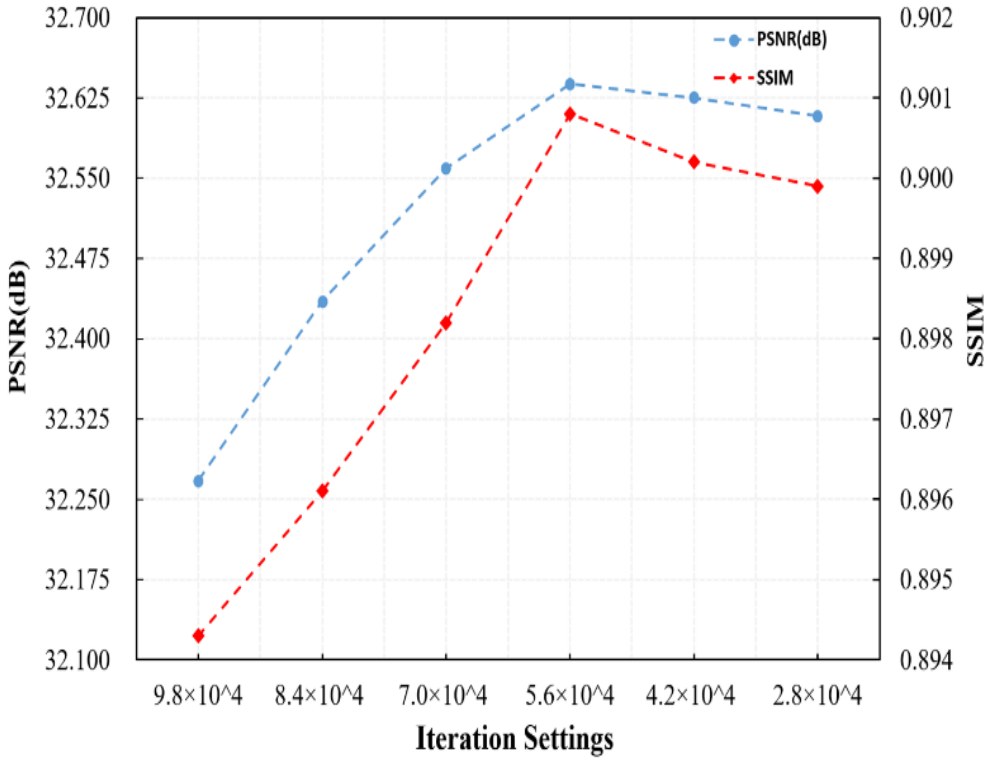

Figure 5. ATG network iteration settings with the BI model on Set5 (4×).

4.4. Ablation Study of Different Modules

To further explore the potential of the proposed method, we comprehensively analyzed our model through ablation experiments. The LNLCN model consists mainly of two parts, namely, the learnable nonlocally residual group (LNLRG) and the adaptive target generator (ATG). We verified their effectiveness by testing them on the Set5 dataset. In addition, we analyzed the benefits of RFP data augmentation for model training, as shown in Table 2.

Table 2. Experimental results of different modules. After 8.9×10^5 iterations, we provide the best PSNR (dB) and SSIM values on Set5.

KERRYPNX	Baseline	R_a	R_b	R_c	R_d	R_e	R_f
Learnable nonlocal contrastive attention (LNLCA)		√			√	√	√
Adaptive target generator (ATG)			√		√		√
RFP data augmentation				√		√	√
Avg. PSNR on Set5 (4×)	32.642	32.692	34.853	38.350	32.686	34.849	38.346
Avg. PSNR on Set5 (3×)	34.746	32.638	34.732	38.307	32.694	34.881	38.356
Avg. PSNR on Set5 (2×)	38.314	32.653	34.794	38.327	32.693	34.874	38.354
Avg. SSIM on Set5 (4×)	0.9003	0.9007	0.9306	0.9624	0.9009	0.9310	0.9627
Avg. SSIM on Set5 (3×)	0.9300	0.9008	0.9309	0.9626	0.9007	0.9308	0.9625
Avg. SSIM on Set5 (2×)	0.9620	0.9003	0.9302	0.9620	0.9009	0.9311	0.9627

4.4.1. Learnable Nonlocal Contrastive Attention (LNLCA)

By testing our model on the Set5 dataset, we verified the importance of the LNLCA module. Specifically, the established baseline network contained 20 LSRAG modules, each of which contained 10 simplified residual blocks, resulting in a network architecture with hundreds of layers, and we added skip connections to each base module. Table 2 shows the performance of the model after 8.9×10^5 training iterations on the Set5 dataset with a downsampling factor of 4, where the baseline model achieves PSNR = 32.642 dB and SSIM = 0.9003. The result for R_a verifies the effectiveness of the learnable nonlocal contrastive attention (LNLCA) module, as the PSNR value increases from the original 32.642 dB to 32.692 dB and the SSIM value increases from the original 0.9003 to 0.9007 compared with the baseline network. Specifically, we combined the LSRAG module in the baseline and our proposed LNLCA module to form a deep feature fusion attention group (DFFAG). The results for R_a are predictable because, compared with the traditional nonlocal attention module adopted by the baseline, we applied the proposed learnable nonlocal contrastive attention to sparsely aggregate the globally relevant information of the image with linear computational complexity, which significantly increased the discriminative learning ability of the network. These experimental results fully prove the superiority of using the LNLCA module.

4.4.2. Adaptive Target Generator (ATG)

We present the results of our ATG training strategy for R_b, R_d, and R_f. Specifically, R_b refers to the use of the ATG training strategy on the basis of the baseline. According to the numerical value, using the ATG training strategy alone slightly reduces the PSNR value of the generated image, whereas the SSIM value is significantly improved. This is to be expected, because the ATG training strategy enables the model to have slightly different outputs, which reduces the erroneous impact of off-target details on the network, thereby restoring the reconstructed images more naturally. As presented in Table 2, both R_d and R_f obtain better SSIM values than R_a and R_e; these metrics indicate that the ATG training strategy has a positive impact on the reconstructed image quality.

4.4.3. RFP Data Augmentation

The results of R_c, R_e, and R_f show that our RFP data augmentation method is beneficial for model training. Specifically, we took the RFP data augmentation method alone as the baseline, and the metric for R_c demonstrates the effectiveness of this method because the PSNR value increases from the original 32.642 dB to 32.653 dB. This is because the adopted RFP data augmentation method achieves data augmentation without changing the image structure, which is beneficial to image reconstruction while ensuring the correlation of pixels. R_e and R_f in Table 2 verify that the RFP data augmentation method can also achieve better results when combined with other modules.

4.5. Comparison with State-of-the-Art Technology (BI)

To further demonstrate the superiority of the proposed LNLCN model, we compared our LNLCN model with nine state-of-the-art SISR models: SRCNN [13], FSRCNN [40], VDSR [14], EDSR [28], DBPN [49], RDN [50], FPAN [51], RCAN [15], ESRT [32], HNCT [34], SAN [20], HAN [52], and NLSN [31]. For comparison, we measured the PSNR and SSIM values on the Y channel of the reconstructed image using a MATLAB [53] function. Table 3 shows the degradation model reconstruction indicators of different methods at scaling factors of two, three, and four. The proposed LNLCN model achieves the best performance at various scaling factors compared to the other SR methods. This is mainly because, compared with the traditional nonlocal modules, we used the learnable nonlocal contrastive attention module to sparsely aggregate image-related information more effectively so that the network could focus on discriminating valid information. In addition, our ATG target training strategy could alleviate the ill-posedness of the super-resolution task by enabling the network to generate slightly different outputs, which is meaningful for obtaining super-

resolution results that are satisfactory in reality and perception, especially in terms of generating pleasing images.

Table 3. Comparison of quantitative results obtained with the BI degradation model. The best and second-best performances are **highlighted** in bold and underlined, respectively.

Method	Scale	Set5 (PSNR/SSIM)	Set14 (PSNR/SSIM)	BSD100 (PSNR/SSIM)	Urban100 (PSNR/SSIM)	Manga109 (PSNR/SSIM)
Bicubic	×2	33.66/0.9299	30.24/0.8688	29.56/0.8431	26.88/0.8403	30.80/0.9339
SRCNN [13]		36.66/0.9542	32.45/0.9067	31.36/0.8879	29.50/0.8946	35.60/0.9663
FSRCNN [40]		37.05/0.9560	32.66/0.9090	31.53/0.8920	29.88/0.9020	36.67/0.9710
VDSR [14]		37.53/0.9590	33.05/0.9130	31.90/0.8960	30.77/0.9140	37.22/0.9750
EDSR [28]		38.11/0.9602	33.92/0.9195	32.32/0.9013	32.93/0.9351	39.10/0.9773
DBPN [49]		38.09/0.9600	33.85/0.9190	32.27/0.9000	32.55/0.9324	38.89/0.9775
RDN [50]		38.24/0.9614	34.01/0.9212	32.34/0.9017	32.89/0.9353	39.18/0.9780
FPAN [51]		38.19/0.9612	33.88/0.9210	32.30/0.9012	32.72/0.9339	39.03/0.9772
RCAN [15]		38.27/0.9614	34.12/0.9216	32.41/0.9027	33.34/0.9384	39.44/0.9786
HNCT [34]		38.08/0.9608	33.65/0.9182	32.22/0.9001	32.22/0.9294	38.87/0.9774
SAN [20]		38.31/0.9620	34.07/0.9213	32.42/0.9028	33.10/0.9370	39.32/0.9792
HAN [52]		38.27/0.9614	34.16/0.9217	32.41/0.9027	33.35/0.9385	39.46/0.9785
NLSN [31]		38.34/0.9618	34.08/**0.9231**	32.43/0.9027	33.42/**0.9394**	39.59/0.9789
LNLCN (Ours)		**38.35/0.9627**	**34.17**/0.9226	**32.46/0.9036**	**33.46**/0.9388	**39.62/0.9798**
Bicubic	×3	30.39/0.8682	27.55/0.7742	27.21/0.7385	24.46/0.7349	26.95/0.8556
SRCNN [13]		30.75/0.9090	29.30/0.8215	28.41/0.7863	26.24/0.7989	30.48/0.9117
FSRCNN [40]		33.18/0.9140	29.37/0.8240	28.53/0.7910	26.43/0.8080	31.10/0.9210
VDSR [14]		33.67/0.9210	29.78/0.8320	28.83/0.7990	27.14/0.8290	32.01/0.9340
EDSR [28]		34.65/0.9280	30.52/0.8462	29.25/0.8093	28.80/0.8653	34.17/0.9476
RDN [50]		34.71/0.9296	30.57/0.8468	29.26/0.8093	28.80/0.8653	34.13/0.9484
FPAN [51]		32.62/0.9291	32.55/0.8467	29.24/0.8090	28.73/0.8642	34.14/0.9481
RCAN [15]		34.74/0.9299	30.65/0.8482	29.32/0.8111	29.09/0.8702	34.44/0.9499
HNCT [34]		34.47/0.9275	30.44/0.8439	29.15/0.8067	28.28/0.8557	33.81/0.9459
ESRT [32]		34.42/0.9268	30.43/0.8433	29.15/0.8063	28.46/0.8574	33.95/0.9455
SAN [20]		34.75/0.9300	30.59/0.8476	29.33/0.8112	28.93/0.8671	34.30/0.9494
HAN [52]		32.75/0.9299	30.67/0.8483	29.32/0.8110	29.10/0.8705	34.48/0.9500
NLSN [31]		34.85/0.9306	30.70/0.8485	29.34/0.8117	**29.25**/0.8726	**34.57**/0.9508
LNLCN (Ours)		**34.87/0.9311**	**30.75/0.8490**	**29.38/0.8119**	29.18/**0.8729**	34.53/**0.9512**
Bicubic	×4	28.42/0.8104	26.00/0.7027	25.96/0.6675	23.14/0.6577	24.89/0.7866
SRCNN [13]		30.48/0.8628	27.50/0.7513	26.90/0.7101	24.52/0.7221	27.58/0.8555
FSRCNN [40]		30.72/0.8660	27.61/0.7550	26.98/0.7150	24.62/0.7280	27.90/0.8610
VDSR [14]		31.35/0.8830	28.02/0.7680	27.29/0.7251	25.18/0.7540	28.83/0.8870
EDSR [28]		32.46/0.8968	28.80/0.7876	27.71/0.7420	26.64/0.8033	31.02/0.9148
DBPN [49]		32.47/0.8980	28.82/0.7860	27.72/0.7400	26.38/0.7946	30.91/0.9137
RDN [50]		32.47/0.8990	28.81/0.7871	27.72/0.7419	26.61/0.8028	31.00/0.9151
FPAN [51]		32.48/0.8984	28.78/0.7867	27.71/0.7412	26.61/0.8025	30.99/0.9144
RCAN [15]		32.63/0.9002	28.87/0.7889	27.77/0.7436	26.82/0.8087	31.22/0.9173
HNCT [34]		32.31/0.8957	28.71/0.7834	27.63/0.7381	26.20/0.7896	30.70/0.9112
ESRT [32]		32.19/0.8947	28.69/0.7833	27.69/0.7379	26.39/0.7962	30.75/0.9100
SAN [20]		32.64/0.9003	28.92/0.7888	27.78/0.7436	26.79/0.8068	31.18/0.9169
HAN [52]		32.64/0.9002	28.90/0.7890	27.80/0.7442	26.85/0.8094	31.42/0.9177
NLSN [31]		32.59/0.9000	28.87/0.7891	27.78/0.7444	**26.96**/0.8109	31.27/0.9184
LNLCN (Ours)		**32.69/0.9006**	**28.95/0.7896**	**27.84/0.7448**	26.94/**0.8112**	**31.44/0.9186**

Visualization Results

In Figure 6, we show the visual quality comparison of SR results with 4×SR under the BI model, from which we can find that most SR models cannot clearly reconstruct the image texture information. Some early models, such as Bicubic, SRCNN, and FSRCNN, cannot even reconstruct the rough outline of the original image. Although some of the more popular methods can reconstruct the general outline of the image, they also suffer from serious artefacts and cannot reconstruct natural textures. Taking "img 8006" as an example,

most existing methods cannot reconstruct the lattice accurately, and there are problems such as blurring and artefacts. Our LNLCN is able to obtain sharper reconstructions with fewer artefacts, which are closer to the original real image. This obvious contrast demonstrates the superiority of our LNLCN model.

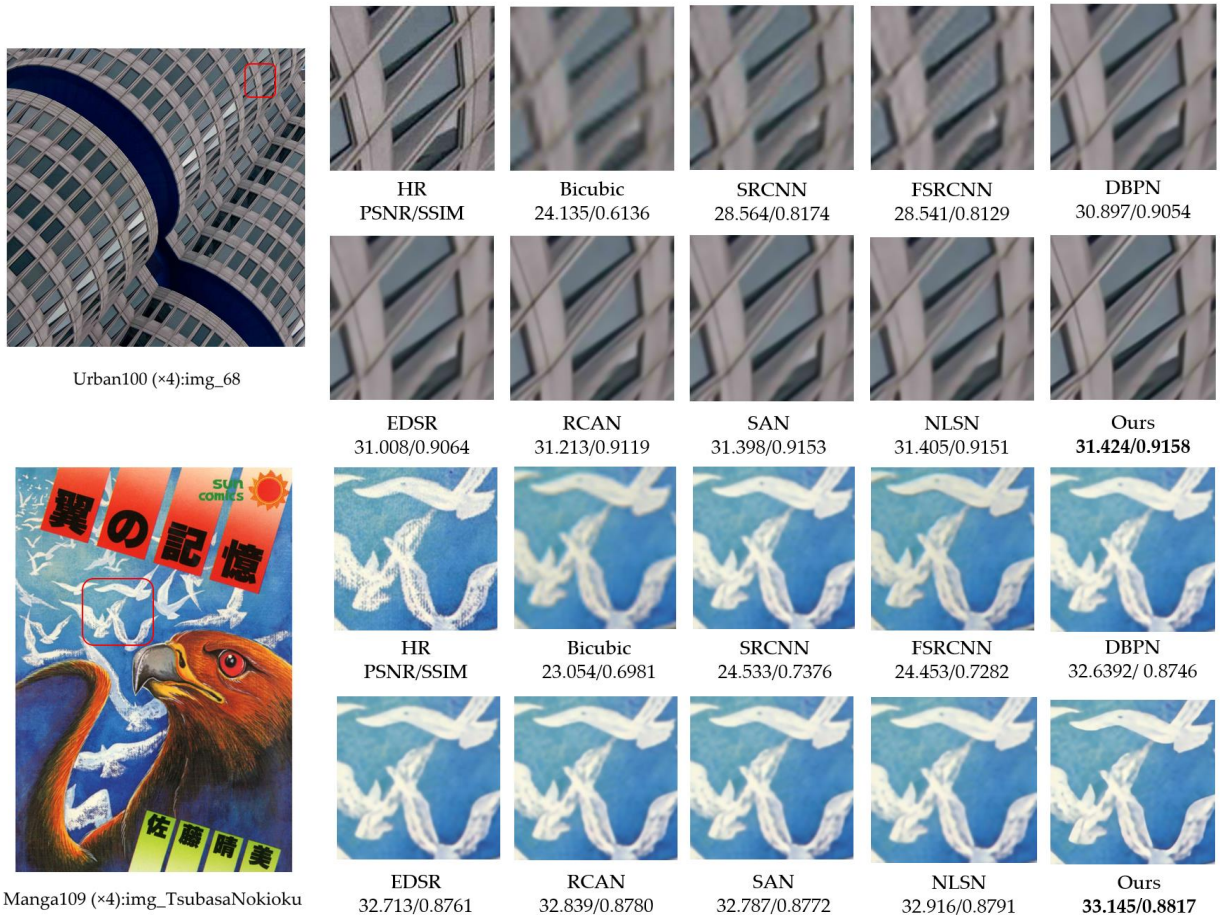

Figure 6. Comparison with visualization results of the BI model on BSD100, Urban100, and Manga109. The best performances are **highlighted** in bold.

4.6. Comparison with State-of-the-Art Technology (BD)

We further compared the LNLCN model with eight state-of-the-art SR models with blur descent (BD) kernels: SRCNN [13], FSRCNN [40], VDSR [14], RCAN [15], IRCNN [54], RDN [50], FPAN [51], SAN [20], and HAN [52]. Table 4 shows the degradation model reconstruction metrics of these models using a scaling factor of three. Our LNLCN model consistently outperforms other models in terms of reconstruction accuracy. Under the condition of blur descent (BD), our proposed learnable nonlocal contrastive attention module can aggregate image-related information more effectively than traditional nonlocal modules, thus enabling the network to better recognize effective information. In addition, the self-similar function setting endows the network with more flexibility. On the Urban100 dataset, the PSNR gain of the LNLCN model is as high as 0.3 dB compared with the RCAN model, which fully verifies the superiority of our LNLCN model.

Visualization Results

In Figure 7, we show the visual quality comparison of SR results with 3 × SR under the BD model. Taking "img_99" as an example, the architectural meshes reconstructed by some methods, such as IRCNN [54] and VDSR [14], have serious artefacts and do not match the reality. The methods with better performance are RDN [50], RCAN [15], and SAN [20], which produce natural overall mesh reconstruction results but still have the

problem of detail texture blurring. Compared with these classic methods, our LNLCN can not only remove image texture blur but also restore high-frequency details, reconstructing relatively sharp edges and clear images. This is because we incorporate learnable nonlocal attention into each DFFAG module, which captures important features which are useful for image texture. IRCNN [54] and RDN [50] are specifically designed to handle image reconstruction tasks with the BD model, and our LNLCN achieves a numerical increase of nearly 1 dB PSNR compared to these models, which fully demonstrates the advantages of the proposed method.

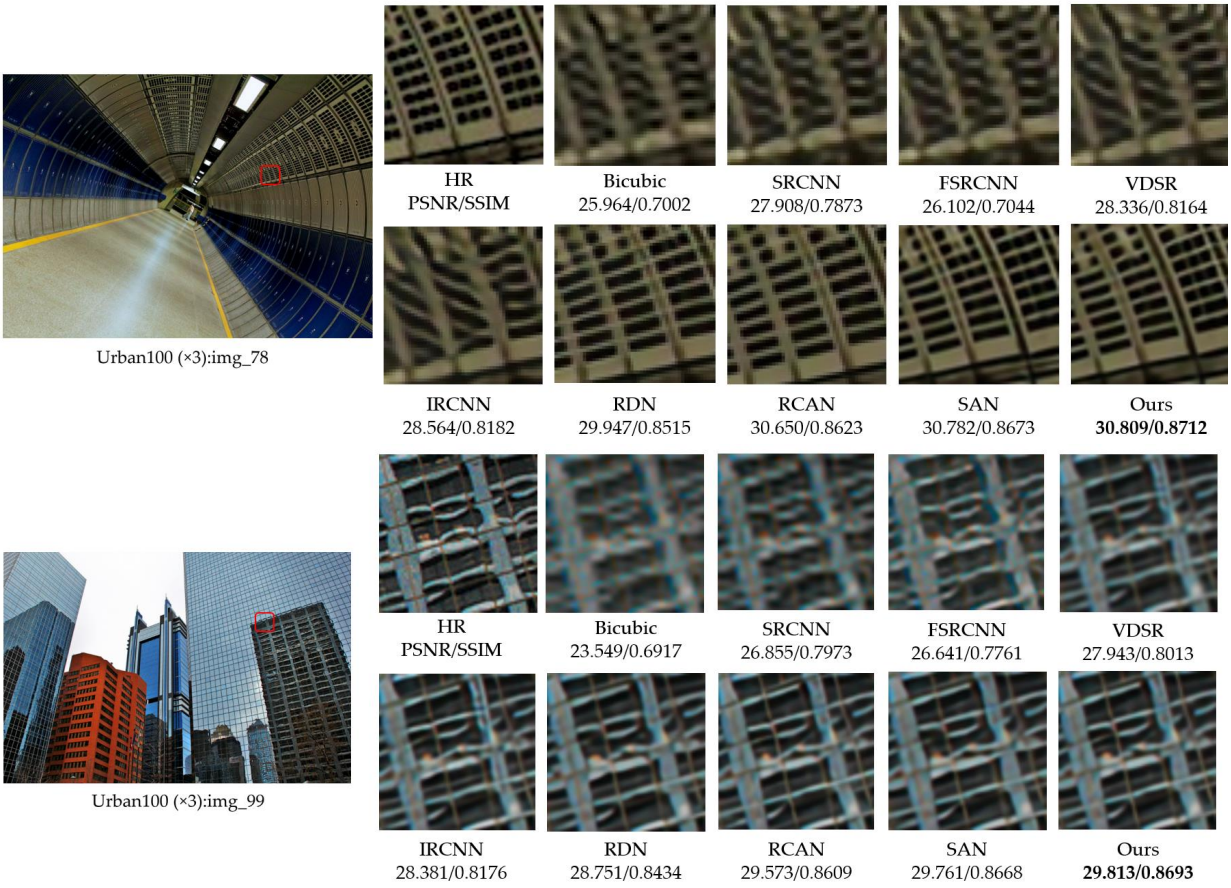

Figure 7. Comparison with visualization results of the BD model on Urban100. The best performances are **highlighted** in bold.

Table 4. Comparison of quantitative results obtained with the BD degradation model. The best and second-best performances are **highlighted** in bold and underlined, respectively.

Method	Scale	Set5 (PSNR/SSIM)	Set14 (PSNR/SSIM)	BSD100 (PSNR/SSIM)	Urban100 (PSNR/SSIM)	Manga109 (PSNR/SSIM)
Bicubic	×3	28.78/0.8308	26.38/0.7271	26.33/0.6918	23.52/0.6862	25.46/0.8149
SRCNN [13]		32.05/0.8944	28.80/0.8074	28.13/0.7736	25.70/0.7770	29.47/0.8924
FSRCNN [40]		26.23/0.8124	24.44/0.7106	24.86/0.6832	22.04/0.6745	23.04/0.7927
VDSR [14]		33.25/0.9150	29.46/0.8244	28.57/0.7893	26.61/0.8136	31.06/0.9234
IRCNN [54]		33.38/0.9182	29.63/0.8281	28.65/0.7922	26.77/0.8154	31.15/0.9245
RDN [50]		34.58/0.9280	30.53/0.8447	29.23/0.8079	28.46/0.8582	33.97/0.9465
RCAN [15]		34.70/0.9288	30.63/0.8462	29.32/0.8093	28.81/0.8645	34.38/0.9483
SAN [20]		34.75/0.9290	30.68/0.8466	29.33/0.8101	28.83/0.8646	34.46/0.9487
HAN [52]		34.76/0.9294	30.70/0.8475	29.34/0.8106	28.99/0.8676	**34.56**/0.9494
LNLCN (Ours)		**34.81/0.9301**	**30.73/0.8482**	**29.35/0.8109**	**29.17/0.8692**	34.55/**0.9506**

5. Discussion

Figure 8 shows the relationship between the reconstruction accuracy of different models and the model parameters. We compared the LNLCA model with 10 state-of-the-art SISR models: VDSR [14], DRCN [26], CARN [55], DBPN [49], RDN [50], EDSR [28], RCAN [15], SAN [20], HAN [52], and NLSN [31]. Among these methods, we observed that some methods, such as DRCN, VDSR, and CARN, contain less computational overhead, but at the cost of reduced reconstruction accuracy. Although some large-scale super-resolution models, such as EDSR and HAN, achieved good results in improving super-resolution performance, the large number of parameters of these models also hinders their use in lightweight applications. In contrast, our proposed LNLCN model has fewer parameters than RDN but achieves higher performance metrics, which means our LNLCN model can achieve good reconstruction results with moderate computational complexity. This is mainly due to the use of Deep Feature Fusion Attention Group (DFFAG) in our LNLCN model, which enables the model to fully utilize the limited LR sample information to achieve more powerful feature representations and thus improve super-resolution performance. Furthermore, our adopted Adaptive Target Generator (ATG) model helps the network explore potential solutions, resulting in sharp outputs. This strategy helps alleviate the ill-posedness of the super-resolution task, enabling the network to generate slightly different outputs, leading to realistically and perceptually pleasing super-resolution results.

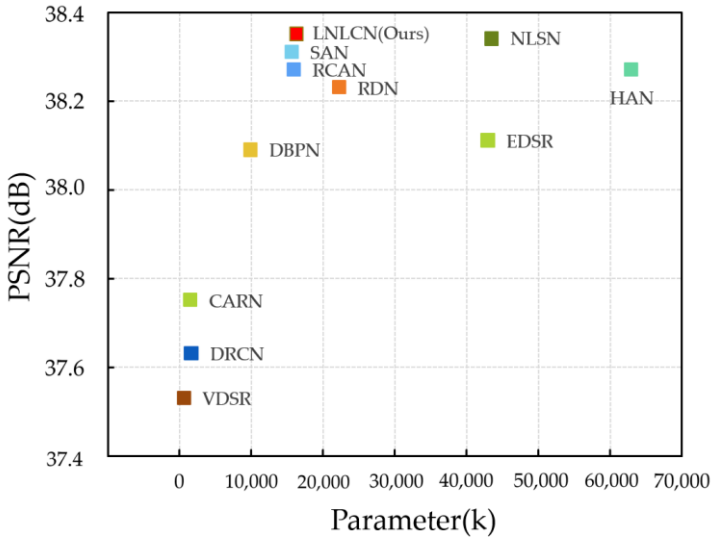

Figure 8. Performance versus model parameters for different models on Set5 (×2).

Table 5 shows the relationship between inference speed and computational cost of different models. We compared the LNLCA model with five state-of-the-art SISR models: VDSR [14], CARN [55], RDN [50], NLSN [31], and HAN [52]. Among these methods, we observed that our proposed LNLCN model performs better in terms of number of parameters, FLOPs, and running time. Compared with other models, the LNLCN model has higher inference speed and lower computational cost.

However, our work also has some limitations. Although the performance on the standard dataset reached a good performance index, the real image has the problem of sparse data samples and unknown degradation kernel. In actual application, the algorithm may not achieve the expected effect. Determining how to simulate the actual image degradation kernel is the key to improving the robustness of the algorithm.

In summary, in future work, we will focus on reconstructing natural and realistic SR images under different degradation conditions. We will use image prior knowledge to improve the accuracy of super-resolution reconstruction in real complex scenes. Specifically, we will use the parser to deal with the degradation factors of different spatial variations,

simulating the situation of super-resolution tasks in real situations, making the model more applicable.

Table 5. Relationship between inference speed and computational cost of different models under 4×.

Method	Params.	FLOPs (4×)	PSNR (Set5 4×)	Running Time
VDSR [14]	670 k	612.6 G	31.35	0.00597 s
CARN [55]	1600 k	90.9 G	32.13	0.00278 s
RDN [50]	22,271 k	1310 G	32.47	0.243 s
NLSN [31]	44,157 k	2956 G	32.59	0.502 s
HAN [52]	64,199 k	3776 G	32.64	0.628 s
LNLCN (Ours)	17,943 k	1175 G	32.69	0.236 s

6. Conclusions

In this paper, we propose a novel learnable nonlocal contrastive network (LNLCN). To reduce the huge computational cost of the nonlocal mechanism, we propose a learnable nonlocal contrastive attention module (LNLCA) and further propose a deep feature fusion module (DFFAG) capable of fusing local adjacency information with nonlocal self-similarity information. Furthermore, we alleviate the ill-posedness of the SR task by introducing an adaptive objective training strategy to seek a potentially optimal solution. Extensive experiments demonstrate that the proposed method has advantages in terms of computational cost and reconstruction effects. The next step will focus on reconstructing natural and realistic SR images under different degradation conditions.

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

Funding: This work was supported in part by Natural Science Foundation of Jiangsu Province under Grant BK20211539, in part by the National Natural Science Foundation of China under Grant U20B2065 and Grant U22B2056, and in part by the Qing Lan Project.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data included in this study are available upon request by contacting the corresponding author.

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

References

1. Zhang, N.; Wang, Y.; Zhang, X.; Xu, D.; Wang, X.; Ben, G.; Zhao, Z.; Li, Z. A multi-degradation aided method for unsupervised remote sensing image super resolution with convolution neural networks. *IEEE Trans. Geosci. Remote Sens.* **2020**, *60*, 5600814. [CrossRef]
2. Dong, X.; Sun, X.; Jia, X.; Xi, Z.; Gao, L.; Zhang, B. Remote Sensing Image Super-Resolution Using Novel Dense-Sampling Networks. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 1618–1633. [CrossRef]
3. Chan, K.C.K.; Zhou, S.; Xu, X.; Loy, C.C. BasicVSR++: Improving video super-resolution with enhanced propagation and alignment. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 5962–5971.
4. Isobe, T.; Jia, X.; Tao, X.; Li, C.; Li, R.; Shi, Y.; Mu, J.; Lu, H.; Tai, Y.W. Look Back and Forth: Video Super-Resolution with Explicit Temporal Difference Modeling. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–20 June 2022; pp. 17411–17420.
5. Pang, Y.; Cao, J.; Wang, J. JCS-Net: Joint Classification and Super-Resolution Network for Small-scale Pedestrian Detection in Surveillance Images. *IEEE Trans. Inf. Forensics Secur.* **2019**, *14*, 3322–3331. [CrossRef]

6. Jiang, J.; Wang, C.; Liu, X.; Ma, J. Deep Learning-based Face Super-resolution: A Survey. *ACM Comput. Surv.* **2021**, *55*, 1–36. [CrossRef]
7. Lugmayr, A.; Danelljan, M.; Gool, L.V.; Timofte, R. Srflow: Learning the super-resolution space with normalizing flow. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 715–732.
8. Zhang, L.; Wu, X. An edge-guided image interpolation algorithm via directional filtering and data fusion. *IEEE Trans. Image Process.* **2006**, *15*, 2226–2238. [CrossRef] [PubMed]
9. Keys, R. Cubic convolution interpolation for digital image processing. In *IEEE Transactions on Acoustics, Speech, and Signal Processing*; IEEE: Piscataway, NJ, USA, 1981; Volume 29, pp. 1153–1160.
10. Wei, Z.; Ma, K.K. Contrast-guided image interpolation. *IEEE Trans. Image Process.* **2013**, *22*, 4271–4285. [CrossRef] [PubMed]
11. Yue, L.; Shen, H.; Li, J.; Yuan, Q.; Zhang, H.; Zhang, L. Image super-resolution: The techniques, applications, and future. *Signal Process.* **2016**, *128*, 389–408. [CrossRef]
12. Zhu, Z.; Guo, F.; Yu, H.; Chen, C. Fast single image super-resolution via self-example learning and sparse representation. *IEEE Trans. Multimed.* **2014**, *16*, 2178–2190. [CrossRef]
13. Dong, C.; Loy, C.C.; He, K.; Tang, X. Learning a Deep Convolutional Network for Image Super-Resolution. In *Computer Vision—ECCV 2014*; Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T., Eds.; Springer: Cham, Switzerland, 2014; Volume 8692, pp. 184–199.
14. Kim, J.; Lee, J.K.; Lee, K.M. Accurate Image Super-Resolution Using Very Deep Convolutional Networks. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 1646–1654.
15. 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.
16. Wang, L.; Wang, Y.; Liang, Z.; Lin, Z.; Yang, J.; An, W.; Guo, Y. Learning parallax attention for stereo image super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019.
17. Yang, F.; Yang, H.; Fu, J.; Lu, H.; Guo, B. Learning texture transformer network for image super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 5791–5800.
18. Wang, T.; Xie, J.; Sun, W.; Yan, Q.; Chen, Q. Dual-Camera Super-Resolution with Aligned Attention Modules. In Proceedings of the IEEE/CVF International Conference on Computer Vision(ICCV), Montreal, QC, Canada, 10–17 October 2021.
19. Magid, S.A.; Zhang, Y.; Wei, D.; Jang, W.D.; Lin, Z.; Fu, Y.; Pfister, H. Dynamic high-pass filtering and multi-spectral attention for image super-resolution. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 4288–4297.
20. Dai, T.; Cai, J.; Zhang, Y.; Xia, S.T.; Zhang, L. Second-order attention network for single image super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, California, CA, USA, 16–20 June 2019; pp. 11065–11074.
21. Xia, B.; Hang, Y.; Tian, Y.; Yang, W.; Liao, Q.; Zhou, J. Efficient Non-local Contrastive Attention for Image Super-resolution. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 22 February–1 March 2022; Volume 36, pp. 2759–2767.
22. Su, J.; Gan, M.; Chen, G.; Yin, J.; Chen, P.C.L. Global Learnable Attention for Single Image Super-Resolution. *arXiv* **2022**, arXiv:2212.01057. [CrossRef] [PubMed]
23. Jo, Y.; Wug Oh, S.; Vajda, P.; Joo Kim, S. Tackling the Ill-Posedness of Super-Resolution through Adaptive Target Generation. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 16231–16240.
24. Yoo, J.; Ahn, N.; Sohn, K.-A. Rethinking data augmentation for image super-resolution: A comprehensive analysis and a new strategy. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–18 June 2020.
25. Shi, W.; Caballero, J.; Huszar, F.; Totz, J.; Aitken, A.P.; Bishop, R.; Rueckert, D.; Wang, Z. Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 1874–1883.
26. Kim, J.; Kwon Lee, J.; Mu Lee, K. Deeply-recursive convolutional network for image super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1637–1645.
27. Tong, T.; Li, G.; Liu, X.; Gao, Q. Image Super-Resolution Using Dense Skip Connections. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 4809–4817.
28. Lim, B.; Son, S.; Kim, H.; Nah, S.; Lee, K.M. Enhanced deep residual networks for single image super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, USA, 21–26 July 2017; pp. 136–144.
29. Ledig, C.; Theis, L.; Huszár, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z.; et al. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 June 2017; pp. 105–114.
30. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 18–21 June 2018.
31. Mei, Y.; Fan, Y.; Zhou, Y. Image super-resolution with non-local sparse attention. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 3517–3526.

32. Lu, Z.; Li, J.; Liu, H.; Huang, C.; Zhang, L.; Zeng, T. Transformer for single image super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 457–466.
33. Chen, R.; Zhang, H.; Liu, J. Multi-Attention augmented network for single image super-resolution. *Pattern Recognit.* **2022**, *122*, 108349. [CrossRef]
34. Fang, J.; Lin, H.; Chen, X.; Zeng, K. A Hybrid Network of CNN and Transformer for Lightweight Image Super-Resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 1103–1112.
35. Hansen, L.K.; Salamon, P. Neural network ensembles. *IEEE Trans. Pattern Anal. Mach. Intell.* **1990**, *12*, 993–1001. [CrossRef]
36. Krogh, A.; Vedelsby, J. Neural network ensembles, cross validation, and active learning. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 1995; pp. 231–238.
37. Guzman-Rivera, A.; Batra, D.; Kohli, P. Multiple choice learning: Learning to produce multiple structured outputs. In *Advances in Neural Information Processing Systems*; Curran Associates Inc.: Lake Tahoe, NV, USA, 2012; pp. 1799–1807.
38. Guzman-Rivera, A.; Kohli, P.; Batra, D.; Rutenbar, A.R. Efficiently enforcing diversity in multi-output structured prediction. In Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics, Reykjavik, Iceland, 22–25 April 2014; pp. 284–292.
39. Guzman-Rivera, A.; Kohli, P.; Glocker, B.; Shotton, J.; Sharp, T.; Fitzgibbon, A.; Izadi, S. Multi-output learning for camera relocalization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, Columbus, OH, USA, 23–28 June 2014.
40. Dong, C.; Loy, C.C.; Tang, X. Accelerating the super-resolution convolutional neural network. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 8–16 October 2016; pp. 391–407.
41. Timofte, R.; Agustsson, E.; Gool, L.V.; Yang, M.H.; Zhang, L.; Limb, B.; Som, S.; Kim, H.; Nah, S.; Lee, K.M.; et al. NTIRE 2017 challenge on single image super-resolution: Methods and results. In Proceedings of the IEEE Conference on CVPRW, Honolulu, HI, USA, 21–26 July 2017; pp. 1110–1121.
42. Bevilacqua, M.; Roumy, A.; Guillemot, C.; Morel, M.L.A. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In Proceedings of the 2012 British Machine Vision Conference, Guildford, UK, 3–7 September 2012. [CrossRef]
43. Zeyde, R.; Elad, M.; Protter, M. On single image scale-up using sparse-representations. In Proceedings of the International Conference on Curves and Surfaces, Avignon, France, 24–30 June 2010; pp. 711–730.
44. Martin, D.; Fowlkes, C.; Tal, D.; Malik, J. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In Proceedings of the ICCV, Vancouver, BC, Canada, 7–14 July 2001; Volume 2, pp. 416–423.
45. Huang, J.B.; Singh, A.; Ahuja, N. Single image super-resolution from transformed self-exemplars. In Proceedings of the IEEE Conference on CVPR, Boston, MA, USA, 7–12 June 2015; pp. 5197–5206.
46. Matsui, Y.; Ito, K.; Aramaki, Y.; Fujimoto, A.; Ogawa, T.; Yamasaki, T.; Aizawa, K. Sketch-based manga retrieval using manga109 dataset. *Multimed. Tools Appl.* **2017**, *76*, 21811–21838. [CrossRef]
47. Kingma, D.P.; Ba, J.L. Adam: A Method for Stochastic Optimization. In Proceedings of the 3rd International Conference for Learning Representations, ICLR 2015, San Diego, CA, USA, 7–9 May 2015.
48. Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; Lerer, A. Automatic differentiation in PyTorch. In Proceedings of the NIPS-W, Long Beach, CA, USA, 4–9 December 2017.
49. Haris, M.; Shakhnarovich, G.; Ukita, N. Deep back-projection networks for super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 1664–1673.
50. Zhang, Y.; Tian, Y.; Kong, Y.; Zhong, B.; Fu, Y. Residual Dense Network for Image Super-Resolution. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 2472–2481.
51. Wu, H.; Gui, J.; Zhang, J.; Kwok, T.J.; Wei, Z. Feedback Pyramid Attention Networks for Single Image Super-Resolution. *IEEE Trans. Circuits Syst. Video Technol.* 2023; *early access*. [CrossRef]
52. Niu, B.; Wen, W.; Ren, W.; Zhang, X.; Yang, L.; Wang, S.; Zhang, K.; Cao, X.; Shen, H. Single Image Super-Resolution via a Holistic Attention Network. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 191–207.
53. The MathWorks, Inc. *MATLAB R2018a*, Version 9.4; The MathWorks, Inc.: Natick, MA, USA, 2018.
54. Zhang, K.; Zuo, W.; Gu, S.; Zhang, L. Learning Deep CNN Denoiser Prior for Image Restoration. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2808–2817.
55. Ahn, N.; Kang, B.; Sohn, K.A. Fast, accurate, and lightweight super-resolution with cascading residual network. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 252–268.