electronics

MDPI

Article

Improving Single-Image Super-Resolution with Dilated Attention

Xinyu Zhang *, Boyuan Cheng, Xiaosong Yang, Zhidong Xiao, Jianjun Zhang and Lihua You *

National Centre for Computer Animation, Faculty of Media and Communication, Talbot Campus, Bournemouth University, Poole BH12 5BB, UK; bcheng@bournemouth.ac.uk (B.C.); xyang@bournemouth.ac.uk (X.Y.); zxiao@bournemouth.ac.uk (Z.X.); jzhang@bournemouth.ac.uk (J.Z.)
* Correspondence: zhangx@bournemouth.ac.uk (X.Z.); lyou@bournemouth.ac.uk (L.Y.)

Abstract: Single-image super-resolution (SISR) techniques have become a vital tool for improving image quality and clarity in the rapidly evolving field of digital imaging. Convolutional neural network (CNN) and transformer-based SISR techniques are very popular. However, CNN-based techniques are not suitable when capturing long-range dependencies, and transformer-based techniques suffer from computational complexity. To tackle these problems, this paper proposes a novel method called dilated attention-based single-image super-resolution (DAIR). It comprises three components: low-level feature extraction, multi-scale dilated transformer block (MDTB), and high-quality image reconstruction. A convolutional layer is used to extract the base features from low-resolution images, which lays the foundation for subsequent processing. Dilated attention is introduced to MDTB to enhance its ability to capture image features at different scales and ensure superior image details and structure recovery. After that, MDTB refines these features to extract multi-scale global attributes and effectively grasps images' long-distance relationships and features across multiple scales. Finally, low-level features obtained from feature extraction and multi-scale global features obtained from MDTB are aggregated to reconstruct high-resolution images. The comparison with existing methods validates the efficacy of the proposed method and demonstrates its advantage in improving image resolution and quality.

Keywords: single-image super-resolution; dilated attention; feature extraction; multi-scale dilated transformer block; image reconstruction

Citation: Zhang, X.; Cheng, B.; Yang, X.; Xiao, Z.; Zhang, J.; You, L. Improving Single-Image Super-Resolution with Dilated Attention. *Electronics* **2024**, *13*, 2281. https://doi.org/10.3390/electronics13122281

Academic Editor: Silvia Liberata Ullo

Received: 7 May 2024
Revised: 6 June 2024
Accepted: 7 June 2024
Published: 11 June 2024

1. Introduction

Super-resolution imaging techniques have long been a pivotal research area in digital image processing. They can be divided into single-image super-resolution (SISR) [1] and multi-image super-resolution (MISR) [2]. Multi-image super-resolution recovers a high-resolution image from multiple low-resolution images of the same object or scene. In contrast, single-image super-resolution reconstructs a high-resolution image from a single low-resolution image. Although multi-image super-resolution is more accurate and robust to noise and other artefacts than single-image super-resolution, it is only applicable to scenarios where two or more low-resolution images of the same object or scene are available. However, there are many situations where only one single low-resolution image is available. For such situations, single-image super-resolution is essential. It is used to enhance image quality, enrich visual experiences, and support advanced analytics in many applications, such as medical imaging, satellite image analysis, video enhancement, security monitoring, communication transmission, virtual production, and the digital multimedia industry. Like most existing research activities that investigate SISR, this paper deals with single-image super-resolution.

Many methods [1,3–7] have been developed for single-image super-resolution. These methods include interpolation-based, reconstruction-based, convolutional neural network-based, and transformer-based ones.

Interpolation-based methods use image interpolation to estimate new pixels with known pixel information. They include nearest-neighbor [8], bilinear [9], and bicubic [10] interpolation. Interpolation-based methods are the simplest and fastest, with low computational cost and real-time performance. However, the images obtained with interpolation-based methods are too smooth and lack high-frequency information. Therefore, they are not applicable to complicated scenes that contain high-frequency details [4,6].

Reconstruction-based methods are also known as regularization-based methods. They reconstruct high-resolution images by inverting the process of down-sampling original high-resolution images, which involves multiple degradation components, such as blurring and warping. Reconstruction-based methods include iterative back-projection [11], projection onto convex sets [12], and maximum a posteriori [13]. They produce better high-resolution images than interpolation-based methods. The disadvantages of reconstruction-based methods include unwanted edges and more sharpness, etc. [14].

Compared to interpolation-based and reconstruction-based methods, learning-based methods significantly improve the performance of SISR. Various learning-based methods have been developed. Among them, convolutional neural network-based and transformer-based methods are the most popular.

Convolutional neural network (CNN)-based methods are capable of handling contextual information efficiently, leading to clearer image reconstruction. For example, IDN gradually extracts abundant and efficient features with distillation blocks [15], and IMDN proposes a distillation module to extract hierarchical features step by step and a fusion module to aggregate the extracted features according to their importance [16]. VDSR uses a very deep convolutional network with 20 weight layers to achieve accuracy improvement [17]. SMSR learns sparse masks and prunes redundant computation to improve efficiency [18]. However, they and other CNNs, such as SRCNN [19], ESPCN [20], and EDSR [21], exhibit limitations in capturing long-range dependencies within an image. Since these methods mainly focus on local features due to their convolutional nature, their ability to integrate contextual information across wider spatial extents is restricted. This can lead to the less effective recovery of detailed and structural information in images, especially when dealing with complex textures or patterns that require understanding broader areas for accurate reconstruction.

Transformer-based methods have been proposed for single-image super-resolution to tackle the incapability of CNNs in capturing long-range dependencies [22]. Transformers [23] were originally proposed for natural language processing [24] and later extended to computer vision [25], including single-image super-resolution [22]. Unlike CNNs, transformer-based methods are capable of modeling long-range dependency. The shortcoming of transformer-based methods is that their computational complexity increases quadratically with increasing spatial resolution, which makes them infeasible in image restoration tasks involving high-resolution images [26].

The above problem can be effectively addressed with dilated attention by using a dilated window with gaps of dilation to increase the receptive field of attention without increasing computational complexity [23]. Talking of this advantage, dilated attention has been introduced to transformers to raise text-processing capacity to one billion tokens [27] and to CNNs to reduce their layers and parameters for image denoising [28].

In this paper, we will introduce dilated attention into transformer-based single-image super-resolution to develop a novel method. It fuses CNN-based low-level feature extraction, dilated attention transformer-based capture and interaction with image features from broader areas and different scales, and high-resolution reconstruction by integrating both low-level and multi-scale global features to achieve more detailed and accurate high-resolution image reconstruction from a single low-resolution image.

Unlike CNNs, our proposed DAIR is built on the dilated attention-based transformer, which significantly expands the model's receptive field without adding computational complexity. In addition, it addresses the transformer's issue of computational complexity,

enabling the transformer to capture long-range dependencies and intricate details across different scales effectively.

The main contributions of this work are summarized below:

- We apply a dilated attention mechanism to SISR tasks to effectively capture image features at different scales and significantly improve detail and structure recovery of images. To the best of our knowledge, single-image super-resolution using a dilated attention-based transformer has not been investigated.
- We fuse low-level features and multi-scale global features to reconstruct images, which ensures high resolution and good quality in terms of the reconstructed images.
- We make a comparison with existing SISR methods to demonstrate the effectiveness and superiority of our proposed DAIR in enhancing image resolution and quality.
- We evaluate the applicability of the proposed method in real-world scenarios, using images with diverse conditions to ensure the method's robustness and generalization capabilities.
- The remaining parts of this paper are organized as follows. Related works are reviewed in Section 2. Our proposed method is introduced in Section 3. The implementation details and experiments are described in Section 4. Conclusions and outlooks are presented in Section 5.

2. Related Works

The work carried out in this paper proposes a new single-image super-resolution method based on a dilated attention-based transformer. In this section, we briefly review the existing work on convolutional neural network-based methods, transformer-based methods, and dilated attention, which are very relevant to the proposed work.

2.1. Convolutional Neural Network-Based Methods

Many convolutional neural network-based methods have been proposed. These methods can be divided into linear network learning, recursive learning, residual learning, dense connection-based learning, progressive learning, multi-scale learning, and attention-based learning.

Linear network learning: With linear networks, multiple convolutional layers are stacked on top of each other to make the input flow sequentially. Linear network learning was pioneered by SRCNN [15], where a convolutional neural network architecture was used to map low-resolution (LR) images to high-resolution (HR) images. Its performance was improved by [29] through a sparse coding-based network (SCN). With SRCNN, the receptive field can be increased by using more convolutional layers to improve the results; however, gradient vanishing/exploding and degradation occur.

Residual learning: Since LR images and HR images mostly share the same information, modeling the residual image between LR and HR is easy. Unlike SRCNN, which learns the desired output directly, residual learning methods, such as [30] AWSRN m [31] and CARN [32], learn the difference between the desired output (HR) and input (LR) to achieve better results but avoid SRCNN's problems of gradient vanishing/exploding and degradation.

Recursive learning: With recursive learning, the same sub-modules (recursive blocks) sharing the same parameters are used repetitively to increase the receptive field without increasing the network parameters. For example, DRCN [33] uses the same convolution as the layer 16 times to obtain a 41 $\times$ 41-sized receptive field. Since too many stacked layers will still cause the gradient vanishing/exploding problem, DRRN [34] bases recursive blocks on residual learning to mitigate the difficulty of training very deep networks. MemNet [35] introduces a memory block, which uses a recursive unit to learn multi-level representations and a gate unit to control the stored amount of previous and current states.

Dense connection-based learning: Not sending the features to the final reconstruction layer, the dense connection-based learning called DenseNet [36] enables each layer to obtain features from all preceding layers, thus creating short paths between most layers to

alleviate the gradient vanishing/exploding problem. SRDenseNet [37] improves DenseNet by using both a dense layer-level connection and a block-level connection. The global-local adjusting dense super-resolution network (GLADSR) links a global-local adjusting module with nested dense connections to use global–local adjusted features more efficiently [38].

Progressive learning: With progressive learning, complicated problems are decomposed into multiple simple tasks to enable a gradual increase in learning difficulty to improve performance and reduce training time. For instance, the sub-band residuals of high-resolution images are gradually reconstructed in LapSRN [39], and the strategy of ordering different difficulties is used to tackle complicated degradation tasks in SRFBN [40].

Multi-scale learning: With multi-scale learning, different features are adaptively extracted at different scales to improve performance through a multi-scale residual block (MSRB) [41]. It is further improved through a multi-scale dense cross block (MDCN) [42] to extract both multi-scale and local features. MADNet enhances multi-scale feature expression through a residual multi-scale module with an attention mechanism [43].

Attention-based learning: The attention mechanism is widely used to improve the performance of SR. Very deep residual channel attention networks (RCANs) introduce a channel-attention-based CNN method to solve SISR [44]. The residual attention module (RAM) combines spatial and channel attention [45]. The residual non-local attention network (RNAN) uses local and global feature attention to create an image restoration network [46]. Residual feature aggregation (RFANet) is proposed to improve spatial attention [47]. The progressive feature fusion network (PFFN) is based on a progressive attention block [48]. The densely residual Laplacian network (DRLN) proposes Laplacian attention [49].

Convolutional neural network-based methods ignore the contextual information outside the local receptive field where convolution is conducted. Thus, they are incapable of extracting such contextual information where a local receptive field is used in convolution and capturing long-range dependencies. This problem can be addressed through the use of transformer-based methods.

2.2. Transformer-Based Methods

Transformers are based on the idea of a self-attention mechanism. They were initially proposed in [50] for English constituency parsing. Since 2020, more and more transformer-based methods have been proposed for SISR.

The texture transformer network for image super-resolution (TTSR) formulates LR images and HR images as transformer queries and keys for SISR [51]. Swin Transformer uses shifted-window-based self-attention to achieve higher efficiency [22]. The linearly assembled pixel-adaptive regression network (LAPAR-A) regresses pixel-adaptive filters, assembles them, and applies them to the bicubic up-sampled image [52]. The lightweight bimodal network integrates a symmetric CNN and a recursive transformer to reduce computational cost and memory consumption [53]. SRFormer builds transformer blocks and layers to capture both global and local features and aggregates the features at different stages [54].

Computational complexity is the main limitation of transformer-based methods. Although various transformer-based methods have been developed, reducing the computational complexity of transformer-based methods is still an important topic.

2.3. Dilated Attention

Dilated attention can be used to increase the receptive field without raising computational complexity. It has been applied in different fields to tackle various problems. In the following, we review some work on dilated attention in computer vision, including SISR.

The dilation networks proposed in [55] introduce dilation to connect elements of a convolution kernel to nonadjacent positions of the previous layer for dense semantic labeling of high-resolution aerial imagery. Dilation is also introduced into deep CNNs to increase the receptive field of SISR but without increasing trainable parameters [56].

The advantages of fewer parameters, less execution time, and better HR results obtained from introducing dilation into deep convolutional neural networks to expand the receptive fields are investigated in [57]. The dilated residual networks with symmetric skip connection (DSNet) combine dilated convolution with symmetric skip connection for image denoising [28]. Dilated-CBAM integrates dilated convolution and a residual network for image classification [58]. The pyramid dilated attention network (PDAN) built on a dilated attention layer addresses the incapability of existing action detection methods in selecting the key temporal information of long videos [59]. The dilated neighborhood attention transformer (DiNAT) [60] incorporates both a localized attention mechanism and dilated neighborhood attention to capture the global context and expand receptive fields for semantic segmentation, etc.

Although various dilated attention-based methods have been developed, we have not found a report on SISR using a dilated attention-based transformer. In this paper, we will investigate it.

3. Proposed Method

Our proposed DAIR is composed of three primary modules: low-level feature extraction, multi-scale global feature extraction, and high-quality image reconstruction. Initially, low-level features are extracted from low-resolution images through a convolutional layer. Subsequently, a dilated transformer block (DTB) is introduced. Equipped with multi-scale null attention and an integrated feed-forward network that includes convolutions, DTB enables the DAIR model to effectively capture and interact with image features across various scales. This design enhances the model's ability to discern and reconstruct image details more accurately. After that, DAIR integrates both low-level and multi-scale global features. Finally, the integrated features are fed into the reconstruction network to achieve good image resolution and quality.

The network architecture of our proposed DAIR is shown in Figure 1. It presents an efficient transformer model with null attention for single-image super-resolution tasks. The flowchart of our proposed DAIR is shown in Figure 2. It can be divided into the following three modules and eight stages.

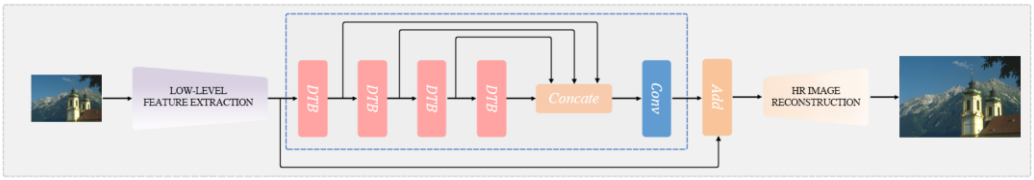

Figure 1. Architecture of the entire DAIR.

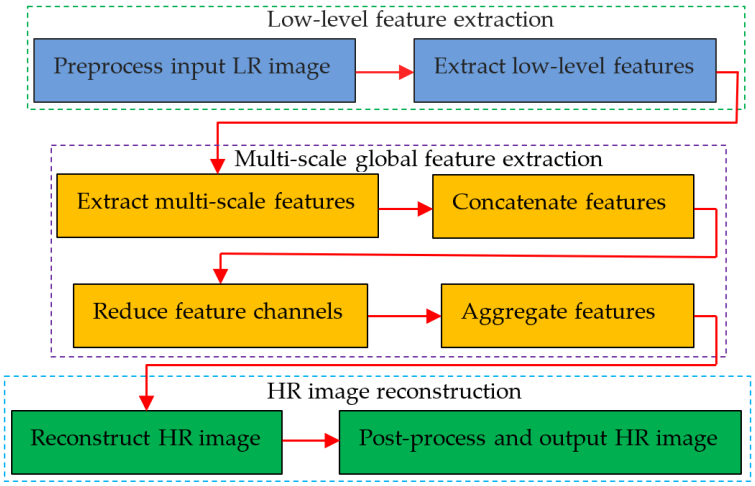

Figure 2. Flowchart of DAIR.

The first module is low-level feature extraction from the input image. It includes the following two stages.

Stage 1: The input low-resolution image is preprocessed through image normalization.

Stage 2: Low-level features are extracted from the normalized image with a 3×3 convolutional layer. More details about low-level feature extraction are given in Section 3.1.

The second module is multi-scale global feature extraction. This task is carried out in a multiple-dilated transformer block (MDTB). It includes the following four stages.

Stage 3: Four dilated transformer blocks (DTBs) are used to model feature long-range dependencies and interact with information at multiple scales. Details of the dilated transformer block are provided in Section 3.2.

Stage 4: The features from the encoder and decoder of the transformer are concatenated in a concatenate layer.

Stage 5: A 1×1 convolution (Conv) layer is used to reduce feature channels.

Stage 6: The low-level features obtained from the first module and the multi-scale global features obtained from Stage 5 are aggregated in an Add layer.

The third module involves high-resolution image reconstruction. It includes the following two stages.

Stage 7: A convolution layer is applied to the aggregated features to reconstruct the high-resolution image.

Stage 8: The reconstructed high-resolution image is post-processed via image denoising, sharpening, and format conversion, and the post-processed high-resolution image is output.

In the following, we first discuss feature extraction in Section 3.1. Then, we elaborate our proposed dilated transformer block (DTB) in Section 3.2. Following that, we introduce high-resolution image reconstruction in Section 3.3. Finally, we provide details of the loss function used in this paper in Section 3.4.

3.1. Feature Extraction

In this subsection, we investigate feature extraction. It includes low-level feature extraction and multi-scale global feature extraction. Given a low-resolution input image $I_l \epsilon \mathbb{R}^{C \times H \times W}$, where $\mathbb{R}^{C \times H \times W}$ is a three-dimensional real space of number of channels, height, and width: C, H, and W correspond to the number of channels, the height, and the width of the input image, respectively, and we first use a convolutional layer $F_{FE}(\cdot)$ of 3×3 to extract the low-level features of the image, which can be mathematically written as follows:

$$F_l = F_{FE}(I_l) \tag{1}$$

where $F_l \epsilon \mathbb{R}^{C \times H \times W}$ represents the low-level features containing the rich detail information of the input image.

After that, we extract multi-scale global features $F_d \epsilon \mathbb{R}^{C \times H \times W}$ from the low-level features F_l with our proposed multi-dilated transformer block, defined as

$$F_d = F_{MDTB}(F_l) \tag{2}$$

where $F_{MDTB}(\cdot)$ indicates a multiple stacked dilated transformer block that includes four transformer blocks with null attention.

The above treatment has the following advantages. On the one hand, the global self-attention in the dilated transformer block can model the long-range contextual dependencies of the features, which compensates for the drawbacks of the limited sensory field of the convolutional coding block. On the other hand, the introduction of cavity convolution in computing attention captures multi-scale features and allows local and sparse feature block interactions in a small range.

3.2. Dilated Transformer Block

Features at different scales contain information about different details in an image. To model feature long-range dependencies and interact with information at multiple scales, we design a dilated transformer block (DTB) shown in Figure 3. It includes a multi-scale dilated attention (MSDA) and a feed-forward network (FFN). The flowchart of DTB is shown in Figure 4 and consists of the following three modules.

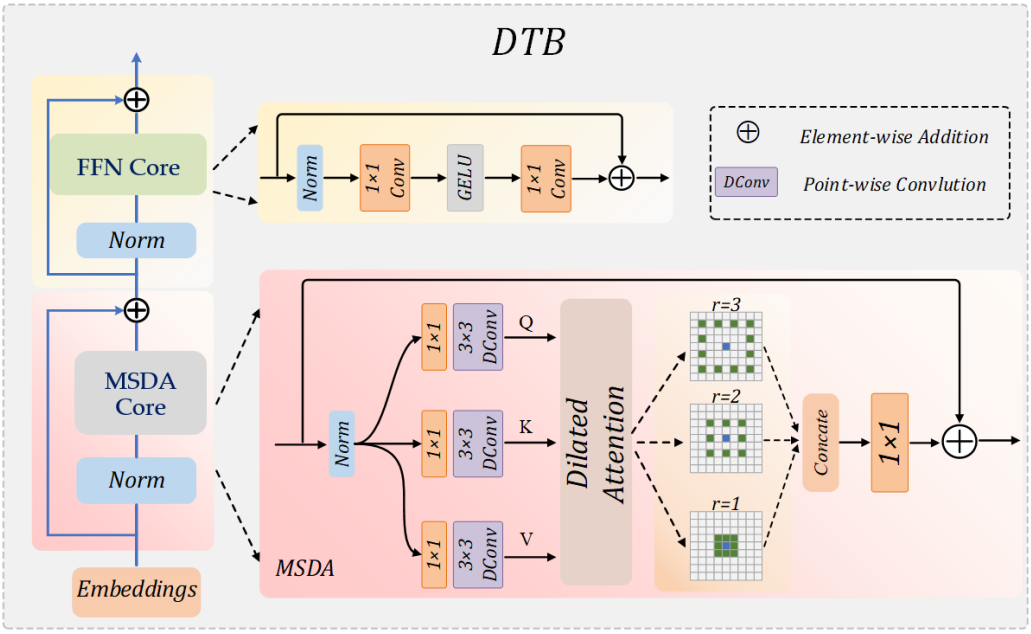

Figure 3. The proposed dilated transformer block (DTB), which consists of three modules: embeddings, multi-scale dilated attention, and feed-forward network. where the blue pixel shows the centre point and the green pixels show the surrounding points of dilated attention.

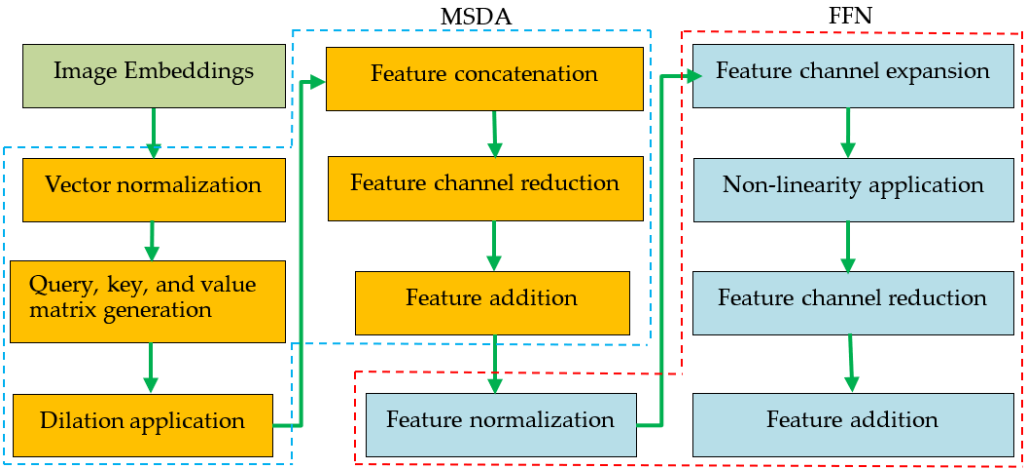

Figure 4. Flowchart of Dilated Transformer Block.

The first module is embeddings, which is used to divide an image into smaller regions denoted as tokens and convert the tokens into vectors.

The second module contains multi-scale dilated attention (MSDA). Its mathematical operation is given in Section 3.2.1, and the process steps are as follows:

Step 1: normalize the vectors obtained from embeddings through a norm layer.

Step 2: generate the query matrix Q_r, key matrix K_r, and value matrix V_r from the normalized vectors, which is achieved by applying the three 1×1 point-wise convolutions

to aggregate pixel-wise cross-channel context followed by the three 3×3 depth-wise convolutions to encode channel-wise spatial context.

Step 3: apply the dilation operation through a dilated attention layer with dilation rates of 1, 2, and 3.

Step 4: concatenate the features obtained from the dilated attention layer.

Step 5: reduce channels with a 1×1 convolution layer.

Step 6: carry out the element-wise addition, which adds the original features before MSDA and the features obtained from Step 5.

The third module is a feed-forward network (FFN). Its mathematical operation is given in Section 3.2.2, and the process steps are as follows:

Step 1: normalize the aggregated features obtained from Step 6.

Step 2: expand the feature channels with the first 1×1 convolution.

Step 3: apply non-linearity via a GELU layer.

Step 4: reduce the feature channels back to the original input dimension with the second 1×1 convolution.

Step 5: conduct the element-wise addition, which adds the features before FFN and the features obtained from Step 4.

3.2.1. Multi-Scale Dilation Attention (MSDA)

Given the input feature $F_l \epsilon \mathbb{R}^{C \times H \times W}$, we first perform layer normalization. Then, three different matrices $Q_r = W^Q W_r^Q F_l$, $K_r = W^K W_r^K F_l$, and $V_r = W^V W_r^V F_l$ are generated where Q_r, K_r, $V_r \epsilon \mathbb{R}^{C \times H \times W}$ are the query, key, and value matrices, W^Q, W^K, and W^V represent the convolutions with kernel size 1×1, W_r^Q, W_r^K, and W_r^V are the depth-separable convolutions with kernel size 3×3, and $r = i$, where $i \in \{1,\ 2,\ 3\}$ denotes the different null rates. After that, we reshape matrix Q_r, K_r, and V_r into smaller shapes: $\widetilde{Q}_l \epsilon \mathbb{R}^{C \times H \times W}$, $\widetilde{K}_l \epsilon \mathbb{R}^{C \times H \times W}$, and $\widetilde{V}_l \epsilon \mathbb{R}^{C \times H \times W}$. The entire computation of attention is defined as follows:

$$F_{attention}^{r}\left(\widetilde{Q}_r, \widetilde{K}_r, \widetilde{V}_r\right) = \widetilde{V}_r \cdot sf\left(\widetilde{K}_r \cdot \widetilde{Q}_r / a\right) \tag{3}$$

where $F_{attention}^{r}(\cdot)$ indicates the entire attention computation, $sf(\cdot)$ represents the SoftMax operation, and a is a section learning variable that controls the size of the dot product of $\widetilde{K}_r$ and $\widetilde{Q}_r$.

Finally, we fuse the features at different scales and sum them with the original input features, defined as follows:

$$F_{attn} = F_{1\times1}(F_{attention}^{r}) + F_l \tag{4}$$

In the above equation, $F_{1\times1}(\cdot)$ indicates 1×1 convolution computation, and F_l and F_{attn} are the input and output feature maps, respectively. Similar to traditional multi-head attention in transformer, we divide the channels into 8 "heads".

3.2.2. Feed-Forward Network (FFN)

To perform feature transformation, we use a feed-forward network for pixel-by-pixel operation. Unlike the conventional multilayer perceptron in transformer, the feed-forward network in this paper consists of two 1×1 convolutions and a GELU activation function, where the first 1×1 convolution is used to extend the number of channels and the second 1×1 convolution is used to reduce the number of channels to the original number of channels. The whole process is defined as follows:

$$F_d = F_2(©F_1(LN(F_{attn}))) + F_{attn} \tag{5}$$

where $LN(\cdot)$ is the layer normalization operation, © stands for the GELU operation, and $F_1(\cdot)$ and $F_2(\cdot)$ are the 1×1 convolutions.

3.3. Image Reconstruction

To perform the image super-resolution task, we aggregate low-level features F_l and multi-scale global features F_d and reconstruct the high-resolution image I_h, defined as follows:

$$I_h = F_{REC}(F_l + F_d) \tag{6}$$

where $F_{REC}(\cdot)$ indicates an image reconstruction module, which consists of sub-pixel convolution.

High-quality images are recovered by up-sampling. Since low-level features typically encompass low-frequency details in an image and the multi-scale global features capture its high-frequency nuances, both low- and high-frequency information is provided to the reconstruction network by utilizing skip connections, which improves the performance and efficiency of our proposed network.

3.4. Loss Function

Mean absolute error is a widely used loss function [43]. In this paper, we use it to optimize the parameters of the network through minimizing the L_1 norm, which is defined as follows:

$$L_1 = \|I_h - I_{rh}\|_1 \tag{7}$$

where $\|\cdot\|_1$ is the L_1 norm operation, I_h is the high-resolution image output from the DAIR network, and I_{rh} is the corresponding real high-resolution image.

4. Experiments

In this section, we first introduce the datasets and evaluation metrics. Then, we give the implementation details and compare reconstructed HR results obtained with our proposed method and existing SISR methods. After that, we present the ablation experiments and our discussion about the experimental results.

4.1. Datasets

We use the DIV2K dataset [61] for training and another five datasets for testing. To choose suitable datasets for evaluating the performance of our proposed model, five datasets, Manga109 [62], BSDS100 [63], Set14 [64], Urban100 [65], and Set5 [66] were selected.

The DIV2K dataset [61] is a high-quality dataset commonly used for single-image super-resolution (SISR) studies. It contains 1000 high-resolution images and provides different low-resolution versions obtained using different down-sampling methods.

The selected five datasets are also widely used in super-resolution studies due to their wide coverage of different scenarios and features. They are very popular for comprehensive evaluations. Manga109 [62] comprises 109 Japanese manga collections. The images in Manga109 possess distinct lines and textures, setting them apart from typical natural images. BSDS100 [63] is part of the Berkeley Segmentation Dataset and contains 100 images. It provides a rich collection of images of natural scenes with various textures and structures. Set14 [64] is a medium-sized benchmark dataset with 14 images. It contains a wide range of scenes, such as buildings, plants, and people. Urban100 [65] is a dataset dedicated to evaluating the performance of super-resolution algorithms in urban scenes. It provides 100 images of buildings, vehicles, road signs, etc. Set5 [66] is a small benchmark dataset consisting of 5 images. It can be used to quickly evaluate the performance of a model.

4.2. Evaluation Metrics

In the field of super-resolution, evaluation metrics are crucial for assessing the performance of a super-resolution model. Two evaluation metrics, PSNR and SSIM, are most widely used in the field of super-resolution. They provide a comprehensive performance evaluation for SISR. Each of them has its own merits.

PSNR is the abbreviation of the Peak Signal-to-Noise Ratio. It is defined by the maximum possible pixel value MAX and the mean squared error MSE between the ground-truth image and the reconstructed image through the following equation:

$$PSNR = 10 \cdot log_{10}\left(\frac{MAX^2}{MSE}\right) \tag{8}$$

where MSE is calculated by

$$MSE = \frac{1}{HWC}\sum_{i=0}^{H-1}\sum_{j=0}^{W-1}\sum_{k=0}^{C-1}\left[I_{gt}(i,\ j,\ k) - I_{re}(i,j,k)\right]^2 \tag{9}$$

where $I_{gt}(i,\ j,\ k)$ is the ground-truth image and $I_{re}(i,\ j,\ k)$ is the reconstructed image.

Due to its simplicity and objectivity, PSNR is a prevalent metric in the realm of image processing and computer vision. By reducing the mean absolute error during the training phase, one can directly enhance the PSNR value.

SSIM is the abbreviation of Structural Similarity Index Measure. It is used to measure the similarity between two images on a perceptual basis and defined by distortion scores on three image features, luminance, contrast, and structure, using the following equation:

$$SSIM = l\left(I_{gt}, I_{re}\right)^{\xi} \cdot c\left(I_{gt}, I_{re}\right)^{\lambda} \cdot s\left(I_{gt}, I_{re}\right)^{\eta} \tag{10}$$

where l, c, and s, respectively, stand for luminance, contrast, and structure between the ground-truth image I_{gt} and the reconstructed image I_{re}, and $\xi > 0$, $\lambda > 0$, and $\eta > 0$ are parameters used to adjust the importance of luminance, contrast, and structure.

In the above equation, $l\left(I_{gt}, I_{re}\right)$, $c\left(I_{gt}, I_{re}\right)$, and $s\left(I_{gt}, I_{re}\right)$ are defined with the following mathematical equations:

$$\begin{aligned} l\left(I_{gt}, I_{re}\right) &= \frac{2\mu_{I_{gt}}\mu_{I_{re}}+C_1}{\mu_{I_{gt}}^2+\mu_{I_{re}}^2+C_1} \\ c\left(I_{gt}, I_{re}\right) &= \frac{2\sigma_{I_{gt}}\sigma_{I_{re}}+C_2}{\sigma_{I_{gt}}^2+\sigma_{I_{re}}^2+C_2} \\ s\left(I_{gt}, I_{re}\right) &= \frac{\sigma_{I_{gt}I_{re}}+C_3}{\sigma_{I_{gt}}\sigma_{I_{re}}+C_3} \end{aligned} \tag{11}$$

Substituting Equation (11) into (10) and setting $\xi = \lambda = \eta = 1$ and $C_3 = C_2/2$, Equation (10) changes to

$$SSIM = \frac{\left(2\mu_{I_{gt}}\mu_{I_{re}} + C_1\right)\left(2\sigma_{I_{gt}I_{re}} + C_2\right)}{\left(\mu_{I_{gt}}^2 + \mu_{I_{re}}^2 + C_1\right)\left(\sigma_{I_{gt}}^2 + \sigma_{I_{re}}^2 + C_2\right)} \tag{12}$$

Unlike PSNR, SSIM considers all three aspects of an image: luminance, contrast, and structure, making it capable of capturing a wide range of image characteristics that are closer to human visual perception. In some cases, an image with a high SSIM value is more visually pleasing than one with a high PSNR value.

4.3. Implementation Details

All experiments were performed on 2 Nvidia RTX 4090 GPUs using the PyTorch (version 1.9) framework. For data enhancement, we used horizontal and vertical flipping. To obtain low-resolution (LR) images, we down-sampled the high-resolution (HR) images using double cubic interpolation. We followed most previous methods in image reconstruction and used Pixel Shuffle to enhance the final rough features into detailed ones. In the training and fine-tuning phases from scratch, we set {patch size, batch size} to {48 × 48, 32} and {64 × 64, 16}, respectively. The training pairs were further enhanced by horizontal flips and random rotations of 90°, 180°, and 270°. We started training with a patch size

of 128 × 128 and a batch size of 64. We introduced 300 K iterations using the AdamW optimizer (β1 = 0.9, β2 = 0.999, weight decay 10^{-4}) and L_1 loss, with an initial learning rate of 3 × 10^{-4} and a gradual decrease to 10^{-6} through cosine annealing [67]. The refinement phase consists of 4 blocks. Our DAIR implements a 4-level encoder–decoder with the number of DTB blocks from level-1 to level-4 as [4, 6, 6, 8], the number of attention heads in MDTA as [1, 2, 4, 8], and the number of channels as [48, 96, 192, 384].

4.4. Comparisons with Existing SISR Methods

Table 1 shows the results comparing DAIR with 12 existing SISR methods, where red indicates the best and blue stands for the second best. Although many of these methods demonstrate excellent performance in the above experiments, DAIR stands out from them.

Table 1. Quantitative comparisons (average PSNR/SSIM) between ours (DAIR) and existing methods across benchmark datasets. Higher PSNR/SSIM number means better quality.

Method	Scale	Params	Set5	BSD100	Set14	Manga109	Urban100
			PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM
IDN [17]	×3	553 K	34.11/0.9253	28.95/0.8013	29.99/0.8354	32.71/0.9381	27.42/0.8359
IMDN [18]	×3	703 K	34.36/0.9270	29.09/0.8046	30.32/0.8417	33.61/0.9445	28.17/0.8519
VDSR [19]	×3	665 K	33.66/0.9213	28.82/0.7976	29.77/0.8314	32.01/0.9310	27.14/0.8279
SMSR [20]	×3	993 K	34.40/0.9270	29.10/0.8050	30.33/0.8412	33.68/0.9445	28.25/0.8536
AWSRN m [31]	×3	1143 K	34.42/0.9275	29.13/0.8059	30.32/0.8419	33.64/0.9450	28.26/0.8545
CARN [32]	×3	1592 K	34.29/0.9255	29.06/0.8034	30.29/0.8407	33.43/0.9427	28.06/0.8493
DRCN [33]	×3	1774 K	33.82/0.9226	28.80/0.7963	29.76/0.8311	32.31/0.9328	27.15/0.8276
DRRN [34]	×3	297 K	34.03/0.9244	28.95/0.8004	29.96/0.8349	32.74/0.9390	27.53/0.8378
MemNet [35]	×3	678 K	34.09/0.9248	28.96/0.8001	30.00/0.8350	32.51/0.9369	27.56/0.8376
GLADSR [38]	×3	821 K	34.41/0.9272	29.08/0.8050	30.37/0.8418	-	28.24/0.8537
MADNet [43]	×3	930 K	34.16/0.9253	28.98/0.8023	30.21/0.8398	-	27.77/0.8439
LAPAR-A [52]	×3	594 K	34.36/0.9267	29.11/0.8054	30.34/0.8421	33.51/0.9441	28.15/0.8523
DAIR	×3	875 K	**34.71/0.9297**	**29.18/0.8084**	**30.68/0.8490**	**33.95/0.9465**	**28.36**/0.8544
IDN [17]	×4	553 K	31.82/0.8903	27.41/0.7297	28.25/0.7730	29.41/0.8942	25.41/0.7632
IMDN [18]	×4	703 K	32.21/0.8948	27.56/0.7353	28.58/0.7811	30.45/0.9075	26.04/0.7838
VDSR [19]	×4	665 K	23.13/0.8838	27.29/0.7251	28.01/0.7674	28.83/0.8809	25.18/0.7524
SMSR [20]	×4	993 K	32.15/0.8944	27.61/0.7366	28.61/0.7818	30.42/0.9074	26.14/0.7871
AWSRN m [31]	×4	1143 K	32.21/0.8954	27.60/**0.7368**	28.65/0.7832	30.56/0.9093	26.15/**0.7884**
CARN [32]	×4	1592 K	32.13/0.8937	27.58/0.7349	28.60/0.7806	30.42/0.9070	26.07/0.7837
DRCN [33]	×4	1774 K	31.53/0.8854	27.23/0.7233	28.02/0.7670	28.98/0.8816	25.14/0.7510
DRRN [34]	×4	297 K	31.68/0.8888	27.38/0.7284	28.21/0.7720	29.46/0.8960	25.44/0.7638
MemNet [35]	×4	678 K	31.74/0.8893	27.4/0.7281	28.26/0.7723	29.42/0.8942	25.50/0.7630
GLADSR [38]	×4	821 K	32.14/0.8940	27.59/0.7361	28.62/0.7813	-	26.12/0.7851
MADNet [43]	×4	930 K	31.95/0.8917	27.47/0.7327	28.44/0.7780	-	25.76/0.7746
LAPAR-A [52]	×4	594 K	32.12/0.8932	27.55/0.7351	28.55/0.7808	30.54/0.9085	26.11/0.7868
DAIR	×4	875 K	**32.62/0.9007**	**27.64**/0.7358	**28.96/0.7904**	**30.77/0.9117**	**26.25**/0.7875

Figure 5 illustrates the visual differences between DAIR and existing SISR methods. The super-resolution images generated by DAIR are superior in texture and detail and visually more appealing.

The quantitative and visual comparisons demonstrate the efficacy of our proposed method and its advantage in improving image resolution and quality. In the following subsection, we further investigate the impacts of FFN and MSDA.

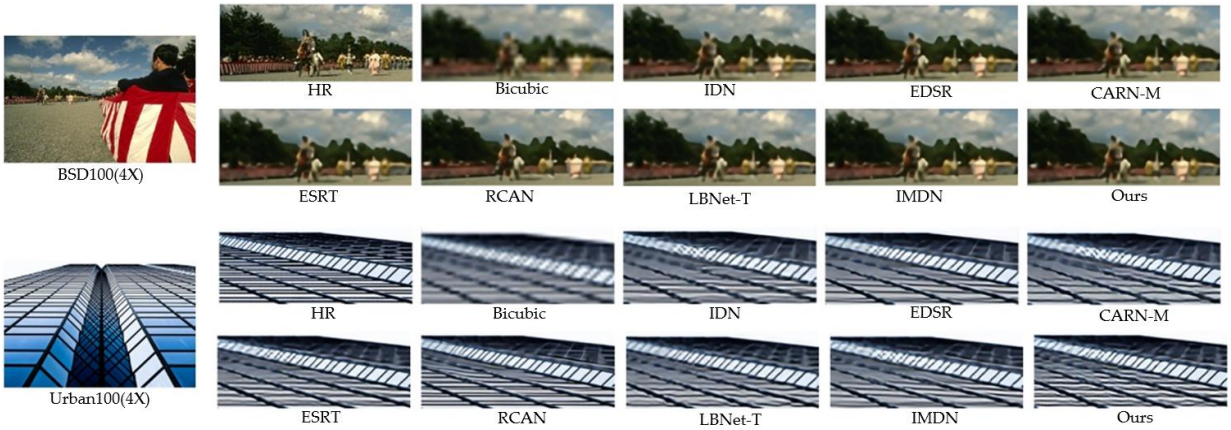

Figure 5. Visual comparisons between DAIR and existing SISR methods.

4.5. Ablation Experiments

Ablation experiments are a commonly used method to evaluate the impact on overall performance by removing or replacing specific components in the model. To further ascertain the significance and efficiency of distinct elements within our model, we carried out ablation studies on (1) the DTB block and (2) low-level feature extraction (LFE) and multi-scale feature extraction (MDTB). The first experiment investigated the impacts of multi-scale dilation attention (MSDA) and a feed-forward network (FFN), and the second experiment examined the influences of LFE and MDTB.

4.5.1. Impacts of MSDA and FFN

For the ablation experiment on MSDA and FFN, we first introduced the experimental setup. Then, we presented the experimental results.

Experimental setup: Three different model variants were designed for the ablation experiment. First, the original model is the complete DAIR model that we have proposed, containing MSDA and FFN. The second variant (w/o FFN) replaced FFN with the standard multilayer perceptrons (MLPs) and kept all other parts the same. The third variant (w/o MSDA) replaced MSDA with the traditional attention mechanism without changing any of the other parts. Every model variant underwent training and testing under identical datasets and conditions to ensure an equitable comparison.

Experimental results: From the experimental results given in Table 2, we can observe the following trends: First, on the Set5 dataset, the original model obtains PSNR 32.62 and SSIM 0.9007. While removing FFN, the performance decreases to PSNR 32.59 and SSIM 0.9004. When MSDA is removed, the performance decreases more significantly to PSNR 32.51 and SSIM 0.8998. Second, on the Set14 dataset, the original model achieves a performance of PSNR 28.96 and SSIM 0.7904. After removing FFN, the performance reduces to PSNR 28.94 and SSIM 0.7900. After removing MSDA, the performance decreases more to PSNR 28.90 and SSIM 0.7899.

Table 2. Ablation experiments for the DTB block where MSDA is replaced with conventional attention and FFN is replaced with MLP.

Methods	Scale	Set5		Set14	
		PSNR	SSIM	PSNR	SSIM
Original model	×4	32.62	0.9007	28.96	0.7904
w/o FFN	×4	32.59	0.9004	28.94	0.7900
w/o MSDA	×4	32.51	0.8998	28.90	0.7899

4.5.2. Influences of LFE and MDTB

For the ablation experiment on LFE and MDTB, we also first discuss the experimental setup. Then, we give the experimental results.

Experimental setup: We designed three different model variants for the ablation experiment. First, the original model is our proposed full DAIR model containing MDTB and LFE. The second variant (w/o LFE) uses MDTB without the initial convolutional layer used for low-level feature extraction. The third variant (w/o MDTB) uses LFE but replaces MDTB with a standard transformer block. All other parts in the second and third variants are kept unchanged. Same as the above, each of the model variants was trained and tested under the same conditions as the benchmark dataset. Their performance was evaluated using the standard metrics PSNR and SSIM.

Experimental results. Table 3 shows the experimental results. For the Set5 dataset, the original model that included LFE and MDTB achieved PSNR 32.62 and SSIM 0.9007. When deleting LFE, the performance decreases to PSNR 32.58 and SSIM 0.9002. When MDTB is replaced with a standard transformer block, the performance reduces much more to PSNR 32.41 and SSIM 0.8990. For the Set14 dataset, the original model obtained PSNR 28.96 and SSIM 0.7904. After removing LFE, they dropped to PSNR 28.95 and SSIM 0.7901. With MDTB replaced by a standard transformer block, they greatly decreased to PSNR 28.83 and SSMI 0.7891.

Table 3. Ablation experiments for LFE and MDTB where LFE is deleted and MDTB is replaced with a standard transformer block.

Methods	Set5		Set14	
	PSNR	**SSIM**	**PSNR**	**SSIM**
Original model	32.62	0.9007	28.96	0.7904
w/o LFE	32.58	0.9002	28.95	0.7901
w/o MDTB	32.41	0.8990	28.83	0.7891

4.6. Discussion

Table 2 shows that both FFN and MSDA contribute to performance improvement. Between them, MSDA made a bigger performance improvement than FFN. This suggests that capturing multi-scale features and long-range contextual information in SISR tasks is more important than feature transformation and fusion. Among the three model variants, the original model achieves the best results, suggesting that including both MSDA and FFN in the DTB block can lead to better performance than including one of them.

Table 3 indicates that both LFE and MDTB improved model performance. Between them, the performance improvement made by MDTB is larger than the performance improvement made by LFE. This justifies the introduction of multi-scale global feature extraction. The original model obtains the best performance among the three model variants. This demonstrates the importance of both low-level feature extraction and multi-scale global feature extraction.

The above ablation experiments validate the effectiveness and importance of DTB, LEF, and MDTB in the DAIR model. MSDA, which plays a vital role in SISR tasks, helps the model capture more abundant and accurate image features. This provides valuable insights for SISR, i.e., more attention should be paid to the extraction and fusion of multi-scale features. Introducing FFN into the DTB block also increases PSNR and SSIM values. Therefore, integrating both MSDA and FFN in the DTB block can maximize the impacts of the DTB block in terms of improving SISR performance.

Compared to the DTB block, the MDTB block has a bigger impact on the model's performance. This is understandable since the DTB block is only a subblock of the MDTB block. Removing LFE results in performance degradation, highlighting the role of low-level feature extraction. The bigger performance improvement of the MDTB block than

LEF shows the importance of multi-scale features and validates the MDTB block in terms of extracting multi-scale features and capturing long-range dependencies. The original model has the highest PSNR and SSIM values. It indicates that low-level feature extraction and multi-scale feature extraction with the MDTB block are essential for achieving high-performance super-resolution.

4.7. Real-World Scenario Evaluation

While the DAIR model has shown promising results on the datasets used in this paper, its performance in real-world scenarios with diverse and unpredictable image conditions remains to be tested. To evaluate the applicability of the proposed method under real-world conditions, we conducted additional experiments using images taken from different environments and under different lighting conditions. Figure 6 shows the results where HR indicates the images from the photos taken, BICUBIC stands for the images reconstructed using bicubic interpolation, and OURS indicates the images reconstructed with our proposed method. The images in Figure 6 highlight the DAIR model's effectiveness in real-world scenarios, confirming its robustness and applicability for practical image enhancement tasks.

Figure 6. Visual results of real-world super-resolution imaging using the DAIR method.

5. Conclusions and Outlooks

In this paper, we developed a new dilated attention-based single-image super-resolution method called DAIR to tackle the incapacity of CNN-based methods in capturing long-range dependencies and the computational complexity of transformer-based methods. It integrates low-level feature extraction, multi-scale global feature extraction across various scales, and high-quality image reconstruction. A specialized convolutional layer extracts low-level image features at the feature extraction stage, which lays the groundwork for subsequent processing. MDTB is the core component of DAIR and consists of four transformer blocks equipped with magnification attention to extract multi-scale global features. The extracted low-level features and multi-scale global features are aggregated to reconstruct high-resolution images. We have compared our proposed method with existing methods to demonstrate its efficacy and advantages in terms of improving image resolution and quality. We also undertook ablation experiments to evaluate the impacts of multi-scale dilation attention and a feed-forward network on network performance.

Despite the fact that the DAIR model has shown promising results on the datasets used in this paper and some real-world applications, in subsequent work, we will further evaluate the performance of our proposed DAIR model in more real-world applications

with diverse and unpredictable image conditions. The aim is to train the model with a more diverse dataset, possibly incorporating images from various sources with different lighting conditions and resolutions. This would make the model more robust and adaptable to real-world scenarios. Computational efficiency can be improved too. Optimization techniques include model pruning and quantization. Pruning removes redundant or less essential weights from a neural network, reducing model size and computational requirements without significantly affecting performance. Quantization reduces the precision of weights and activations from floating-point representations to lower bit-width representations (e.g., 8-bit integers), thus reducing the model size and speeding up the computation.

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

Funding: This research received no external funding.

Data Availability Statement: The data presented in this study are openly available from http://www.manga109.org/en/ (Manga109), https://paperswithcode.com/dataset/bsd (BSDS100), https://github.com/jbhuang0604/SelfExSR?tab=readme-ov-file (Set14, Urban100, Set5), all the above datasets were accessed on 1 February 2023.

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

References

1. Chen, H.; He, X.; Qing, L.; Wu, Y.; Ren, C.; Sheriff, R.E.; Zhu, C. Real-world single image super-resolution: A brief review. *Inf. Fusion.* **2022**, *79*, 124–145. [CrossRef]
2. Park, S.C.; Park, M.K.; Kang, M.G. Super-resolution image reconstruction: A technical overview. *IEEE Signal Process Mag.* **2003**, *20*, 21–36. [CrossRef]
3. Li, J.; Pei, Z.; Zeng, T. From beginner to master: A survey for deep learning-based single-image super-resolution. *arXiv* **2021**, arXiv:2109.14335.
4. Yu, M.; Shi, J.; Xue, C.; Hao, X.; Yan, G. A review of single image super-resolution reconstruction based on deep learning. *Multimed. Tools Appl.* **2023**, *83*, 55921–55962. [CrossRef]
5. Chauhan, K.; Patel, S.N.; Kumhar, K.; Bhatia, J.; Tanwar, S.; Davidson, I.E.; Mazibuko, T.F.; Sharma, R. Deep learning-based single-image super-resolution: A comprehensive review. *IEEE Access* **2023**, *11*, 21811–21830. [CrossRef]
6. Al-Mekhlafi, H.; Liu, S. Single image super-resolution: A comprehensive review and recent insight. *Front. Comput. Sci.* **2024**, *18*, 181702. [CrossRef]
7. Li, J.; Pei, Z.; Li, W.; Gao, G.; Wang, L.; Wang, Y.; Zeng, T. A systematic survey of deep learning-based single-image super-resolution. *ACM Comput. Surv.* **2024**, accepted. [CrossRef]
8. Wang, Y.; Wan, W.; Wang, R.; Zhou, X. An improved interpolation algorithm using nearest neighbor from VTK. In Proceedings of the 2010 International Conference on Audio, Language and Image Processing, Shanghai, China, 23–25 November 2010; pp. 1062–1065.
9. Parsania, P.; Virparia, D. A review: Image interpolation techniques for image scaling. *Int. J. Innov. Res. Comput. Commun. Eng.* **2014**, *2*, 7409–7414. [CrossRef]
10. Gavade, A.B.; Sane, P. Super resolution image reconstruction by using bicubic interpolation. In Proceedings of the National Conference on Advanced Technologies in Electrical and Electronic Systems, Pune, India, 19–20 February 2014; pp. 201–209.
11. Irani, M.; Peleg, S. Improving resolution by image registration. *Graph. Models Image Process.* **1991**, *53*, 231–239. [CrossRef]
12. Stark, H.; Oskoui, P. High-resolution image recovery from image-plane arrays, using convex projections. *J. Opt. Soc. Am. A* **1989**, *6*, 1715–1726. [CrossRef]
13. Schultz, R.R.; Stevenson, R.L. Extraction of high-resolution frames from video sequences. *IEEE Trans. Image Process.* **1996**, *5*, 996–1011. [CrossRef] [PubMed]
14. Lepcha, D.C.; Goyal, B.; Dogra, A.; Goyal, V. Image super-resolution: A comprehensive review, latest trends, challenges and applications. *Inf. Fusion.* **2023**, *91*, 230–260. [CrossRef]
15. Hui, Z.; Wang, X.; Gao, X. Fast and accurate single image super-resolution via information distillation network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 723–731.
16. Hui, Z.; Gao, X.; Yang, Y.; Wang, X. Lightweight image super-resolution with information multi-distillation network. In Proceedings of the 27th ACM International Conference on Multimedia, Nice, France, 21–25 October 2019; pp. 2024–2032.

17. 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, Las Vegas, NV, USA, 27–30 June 2016; pp. 1646–1654.
18. Wang, L.; Dong, X.; Wang, Y.; Ying, X.; Lin, Z.; An, W.; Guo, Y. Exploring sparsity in image super-resolution for efficient inference. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 4915–4924.
19. Dong, C.; Loy, C.C.; He, K.; Tang, X. Image super-resolution using deep convolutional networks. *arXiv* **2015**, arXiv:1501.00092v3. [CrossRef] [PubMed]
20. Shi, W.; Caballero, J.; Ferenc Huszár, 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, Las Vegas, NV, USA, 27–30 June 2016; pp. 1874–1883.
21. Lim, B.; Son, S.; Kim, H.; Nah, S.; Lee, K.M. Enhanced deep residual networks for single image super-resolution. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 July 2017; pp. 1132–1140.
22. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 9992–10002.
23. Lin, T.; Wang, Y.; Liu, X.; Qiu, X. A survey of transformers. *AI Open* **2022**, *3*, 111–132. [CrossRef]
24. Sutskever, I.; Vinyals, O.; Le, Q.V. Sequence to sequence learning with neural networks. In Proceedings of the Advances in Neural Information Processing Systems 27, Montreal, QC, Canada, 8–13 December 2014; pp. 3104–3112.
25. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. In Proceedings of the 2021 International Conference on Learning Representations, Virtual Event, 3–7 May 2021; pp. 1–21.
26. Zamir, S.W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F.S.; Yang, M.-H. Restormer: Efficient transformer for high-resolution image restoration. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 19–24 June 2022; pp. 5718–5729.
27. Ding, J.; Ma, S.; Dong, L.; Zhang, X.; Huang, S.; Wang, W.; Zheng, N.; Wei, F. LONGNET: Scaling transformers to 1,000,000,000 tokens. *arXiv* **2023**, arXiv:2307.02486v2.
28. Peng, Y.; Zhang, L.; Liu, S.; Wu, X.; Zhang, Y.; Wang, X. Dilated residual networks with symmetric skip connection for image denoising. *Neurocomputing* **2019**, *345*, 67–76. [CrossRef]
29. Wang, Z.; Liu, D.; Yang, J.; Han, W.; Huang, T. Deep networks for image super-resolution with sparse prior. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 370–378.
30. 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.
31. Wang, C.; Li, Z.; Shi, J. Lightweight image super-resolution with adaptive weighted learning network. *arXiv* **2019**, arXiv:1904.02358.
32. Ahn, N.; Kang, B.; Kyung, K.-A. Fast, accurate, and lightweight super-resolution with cascading residual network. In Proceedings of the 2018 European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 1–17.
33. Kim, J.; Lee, J.K.; Lee, K.M. Deeply-recursive convolutional network for image super-resolution. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1637–1645.
34. Tai, Y.; Yang, J.; Liu, X. Image super-resolution via deep recursive residual network. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2790–2798.
35. Tai, Y.; Yang, J.; Liu, X.; Xu, C. MemNet: A persistent memory network for image restoration. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 4549–4557.
36. Huang, G.; Liu, Z.; van der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
37. Tong, T.; Li, G.; Liu, X.; Gao, Q. Image super-resolution using dense skip connections. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 4799–4807.
38. Zhang, X.; Gao, P.; Liu, S.; Zhao, K.; Li, G.; Yin, L.; Chen, C.W. Accurate and efficient image super-resolution via global-local adjusting dense network. *IEEE Trans. Multimedia* **2021**, *23*, 1924–1937. [CrossRef]
39. Lai, W.-S.; Huang, J.-B.; Ahuja, N.; Yang, M.-H. Deep Laplacian pyramid networks for fast and accurate super-resolution. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 5835–5843.
40. Li, Z.; Yang, J.; Liu, Z.; Yang, X.; Jeon, G.; Wu, W. Feedback network for image super-resolution. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 3867–3876.
41. Li, J.; Fang, F.; Mei, K.; Zhang, G. Multi-scale residual network for image super-resolution. *Lect. Notes Comput. Sci.* **2018**, *11212*, 527–542.
42. Li, J.; Fang, F.; Li, J.; Mei, K.; Zhang, G. MDCN: Multi-scale dense cross network for image super-resolution. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *31*, 2547–2561. [CrossRef]

43. Lan, R.; Sun, L.; Liu, Z.; Lu, H.; Pang, C.; Luo, X. MADnet: A fast and lightweight network for single-image super resolution. *IEEE Trans. Cybern.* **2021**, *51*, 1443–1453. [CrossRef]
44. 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 15th European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 294–310.
45. Kim, J.H.; Choi, J.H.; Cheon, M.; Lee, J.S. Ram: Residual attention module for single image super-resolution. *arXiv* **2018**, arXiv:1811.12043v1.
46. Zhang, Y.; Li, K.; Li, K.; Zhong, B.; Fu, Y. Residual non-local attention networks for image restoration. *arXiv* **2019**, arXiv:1903.10082v1.
47. Liu, J.; Zhang, W.; Tang, Y.; Tang, J.; Wu, G. Residual feature aggregation network for image super-resolution. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 2356–2365.
48. Zhang, D.; Li, C.; Xie, N.; Wang, G.; Shao, J. PFFN: Progressive feature fusion network for lightweight image super-resolution. In Proceedings of the 29th ACM International Conference on Multimedia, Chengdu, China, 20–24 October 2021; pp. 3682–3690.
49. Anwar, S.; Barnes, N. Densely residual Laplacian super-resolution. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 1192–1204. [CrossRef] [PubMed]
50. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
51. Yang, F.; Yang, H.; Fu, J.; Lu, H.; Guo, B. Learning texture transformer network for image super-resolution. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 5790–5799.
52. Li, W.; Zhou, K.; Qi, L.; Jiang, N.; Lu, J.; Jia, J. LAPAR: Linearly-assembled pixel-adaptive regression network for single image super-resolution and beyond. In Proceedings of the 34th Conference on Neural Information Processing Systems, Virtual Event, 6–12 December 2020; pp. 1–13.
53. Gao, G.; Wang, Z.; Li, J.; Li, W.; Yu, Y.; Zeng, T. Lightweight bimodal network for single-image super-resolution via symmetric CNN and recursive transformer. In Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, Vienna, Austria, 23–29 July 2022; pp. 913–919.
54. Mehri, A.; Behjati, P.; Carpio, D.; Sappa, A.D. SRFormer: Efficient yet powerful transformer network for single image super resolution. *IEEE Access* **2023**, *11*, 121457–121469. [CrossRef]
55. Sherrah, J. Fully convolutional networks for dense semantic labelling of high-resolution aerial imagery. *arXiv* **2016**, arXiv:1606.02585.
56. Huang, Z.; Wang, L.; Meng, G.; Pan, C. Image super-resolution via deep dilated convolutional networks. In Proceedings of the 2017 IEEE International Conference on Image Processing, Beijing, China, 17–20 September 2017; pp. 953–957.
57. Shamsolmoali, P.; Li, X.; Wang, R. Single image resolution enhancement by efficient dilated densely connected residual network. *Signal Process. Image Commun.* **2019**, *79*, 13–23. [CrossRef]
58. Yang, J.; Jiang, J. Dilated-CBAM: An efficient attention network with dilated convolution. In Proceedings of the 2021 IEEE International Conference on Unmanned Systems, Nanjing, China, 30 October–1 November 2021; pp. 11–15.
59. Dai, R.; Das, S.; Minciullo, L.; Garattoni, L.; Francesca, G.; Bremond, F. PDAN: Pyramid dilated attention network for action detection. In Proceedings of the 2021 IEEE Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 5–9 January 2021; pp. 2969–2978.
60. Hassani, A.; Shi, H. Dilated neighborhood attention transformer. *arXiv* **2022**, arXiv:2209.15001v3.
61. Agustsson, E.; Timofte, R. NTIRE2017 Challenge on single image super-resolution: Dataset and study. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 July 2017; pp. 1122–1131.
62. Matsui, Y.; Ito, K.; Aramaki, Y.; Fujimoto, A.; Ogawa, T.; Yamasaki, T.; Aizawa, K. Sketch-based manga retrieval using manga 109 dataset. *arXiv* **2017**, arXiv:1510.04389v1.
63. 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 Eighth IEEE International Conference on Computer Vision, Vancouver, BC, Canada, 7–14 July 2001; pp. 416–423.
64. Zeyde, R.; Elad, M.; Protter, M. On single image scale-up using sparse-representations. *Lect. Notes Comput. Sci.* **2010**, *6920*, 1–20.
65. Huang, J.-B.; Singh, A.; Ahuja, N. Single image super-resolution from transformed self-exemplars. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 5197–5206.
66. Marco, B.; Roumy, A.; Guillemot, C.M.; Alberi-Morel, M.-L. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In Proceedings of the British Machine Vision Conference, Surrey, UK, 3–7 September 2012; pp. 1–10.
67. Loshchilov, I.; Hutter, F. SGDR: Stochastic gradient descent with warm restarts. In Proceedings of the 2017 International Conference on Learning Representations, Toulon, France, 24–26 April 2017; pp. 1–16.