Article

Perceptual Metric Guided Deep Attention Network for Single Image Super-Resolution

Yubao Sun, Yuyang Shi, Ying Yang and Wangping Zhou *

Jiangsu Key Laboratory of Big Data Analysis Technology, Jiangsu Collaborative Innovation Center of Atmospheric Environment and Equipment Technology, Nanjing University of Information Science and Technology, Nanjing 210044, China; sunyb@nuist.edu.cn (Y.S.); yuyang.shi@nuist.edu.cn (Y.S.); yingyang@nuist.edu.cn (Y.Y.)

* Correspondence: wpzhou@nuist.edu.cn; Tel.: +86-189-5199-7253

Received: 16 June 2020; Accepted: 13 July 2020; Published: 15 July 2020

Abstract: Deep learning has been widely applied to image super-resolution (SR) tasks and has achieved superior performance over traditional methods due to its excellent feature learning capabilities. However, most of these deep learning-based methods require training image sets to pre-train SR network parameters. In this paper, we propose a new single image SR network without the need of any pre-training. The proposed network is optimized to achieve the SR reconstruction only from a low resolution observation rather than training image sets, and it focuses on improving the visual quality of reconstructed images. Specifically, we designed an attention-based decoder-encoder network for predicting the SR reconstruction, in which a residual spatial attention (RSA) unit is deployed in each layer of decoder to capture key information. Moreover, we adopt the perceptual metric consisting of L1 metric and multi-scale structural similarity (MSSSIM) metric to learn the network parameters. Different than the conventional MSE (mean squared error) metric, the perceptual metric coincides well with perceptual characteristics of the human visual system. Under the guidance of the perceptual metric, the RSA units are capable of predicting the visually sensitive areas at different scales. The proposed network can thus pay more attention to these areas for preserving visual informative structures at multiple scales. The experimental results on the Set5 and Set14 image set demonstrate that the combination of Perceptual metric and RSA units can significantly improve the reconstruction quality. In terms of PSNR and structural similarity (SSIM) values, the proposed method achieves better reconstruction results than the related works, and it is even comparable to some pre-trained networks.

Keywords: super-resolution; generator network; residual attention; perceptual metric

1. Introduction

Single Image Super-resolution (SISR) is designed to generate a high resolution (HR) image from a single low resolution (LR) image, which has been used for a variety of vision related tasks, such as remote sensing and imaging [1], medical imaging [2], and image enhancement. A variety of SISR methods have been proposed, including prediction-based methods [3], edge-based methods [4], statistical methods [5], patch-based methods [6], sparse representation methods [7], etc. These methods rely primarily on some pre-defined prior models to represent the underlying HR image, which are recognized as model-driven reconstruction methods. With the rapid development of deep learning technology, deep networks, especially convolution neural networks (CNNs), have been widely used for image generation [8] and super-resolution (SR) reconstruction [9], due to their superior performance over model-driven methods. Their main ideas are to train the deep networks to learn the inverse reconstruction mapping from LR images to HR images [10]. Although deep learning-based methods have good reconstruction quality, they are data driven, and a large set of pair-wised LR and HR

images are required for network pre-training, which limits their applicability in practical scenarios. In addition, when the structural features of the testing images are inconsistent with the training images, the reconstruction quality may be degenerated.

Recently, Dmitry Ulyanov et al. proposed the Deep Image Prior (DIP) model [11], which uses a randomly initialized network as a parameterized representation of an image. DIP does not require a large amount of images for network pre-training. Based on the observation model of a degraded image, the DIP model reconstructs the original image from the degraded observation by iteratively optimizing the maximum likelihood estimation of the network parameters. Nevertheless, there is still a performance gap between DIP and the pre-trained networks. The network architecture and the loss function are two important factors that affect the reconstruction performance. In DIP, an hourglass-like network with MSE (mean squared error) loss is used. However, it is generally believed that MSE loss, namely l_2 loss, is inconsistent with perceived characteristics of human vision system [12,13]. Therefore, minimizing the MSE loss cannot necessarily maximize the visual quality of the reconstructed image [12,13]. The network architecture should be further improved to capture visually sensitive features.

In order to cope with these problems, we propose a perceptual metric guided deep attention network (abbreviated as PM-DAN) for predicting SR reconstruction. The overall flowchart of our method is shown in Figure 1. The main strength of our work is to design an advisable network architecture for improving visual quality of the reconstructed image. Specifically, the attention-based encoder-decoder network is constructed for generating the unknown HR image, in which residual spatial attention (RSA) units are deployed in each layer of encoder. Moreover, we adopt the perceptual metric, namely the hybrid of l_1 metric and multi-scale structural similarity (MSSSIM) metric [12], to guide the network learning, which can promote RSA units to attach importance to visually sensitive structures, thereby improving visual quality of reconstructed HR images. The experimental results show that the proposed model can outperform the DIP model in terms of both SSIM and PSNR and is even comparable to some pre-trained networks.

Figure 1. The flowchart of the proposed perceptual metric guided deep attention network (PM-DAN) for single image super-resolution (SISR).

2. Related Work

In recent years, deep convolutional neural networks (CNN) have been widely used for image generation and SISR, and have effectively improved the quality of reconstructed images [14]. Dong et al. first exploited a convolutional neural network named SRCNN to perform SISR reconstruction [15]. In order to enrich the network capacity, some follow-up methods, such as VDSR [16] and IRCNN [17], continued to increase the network depth by stacking more convolutional layers. However, deeper networks require more image samples to train well, while bringing performance improvements. Jiwon et al. proposed a deeply-recursive convolutional network (DRCN) for SR [18] reconstruction, in which a recursive layer (up to 16 recursions) is used to increase the network depth without introducing new network parameters. These methods need to first upscale the LR image into an interpolated image with the same resolution as the HR image, and then feed it into the SR network. However, some useful information may be lost during the interpolation operation, and convolution operations in the HR space increase computational complexity.

Some other studies advocated to learn the end to end mapping from the original LR to the HR image directly. Reference [19] used multiple deconvolution layers to upscale the resolution of feature maps until to be same as that of the HR image. In Reference [20], Shi et al. proposed an efficient sub-pixel convolution layer for upscaling image resolution. Then, EDSR [21] and SRResNet [22] employed sub-pixel convolutions to increase the resolution of the network output, where the residual block is also used to learn the reconstruction mapping. In order to capture multi-scale structures of images, Reference [23] proposed a laplacian pyramid SR network (LapSRN), which can reconstruct the sub-band residuals of the high-resolution image at multiple pyramid scales progressively, and its reconstruction performance is better than SRCNN [15], VDSR [16], and DRCN [18]. Moreover, LapSRN can produce multi-scale SR images (e.g., $2\times$, $4\times$ and $8\times$) with a single feed-forward pass.

All the above-mentioned methods learn SR mapping in supervised ways. Although they can produce promising reconstruction results, a large set of image pairs, consisting of the LR images and the corresponding ground-truth HR images, are required to pre-train the network parameters, which limits the applicability of these methods in practical scenarios. In some practical problems, the real HR images are not easily collected or even unavailable [24]. At the same time, if statistical characteristics of test images deviate significantly from training images, the reconstruction quality will be degraded [25]. The recent work DIP is a parametric network for image representation [11] without the need of pre-training on a large image set. The motivation behind DIP is that the convolutional network itself acts as a good priori of image structures, and the network parameters can be optimized to represent the single instance under a given observation model. DIP provides a new approach for single image SR. We go further upon this model and propose an unsupervised single-image SR network. Different from maximizing the PSNR metric in Reference [11], the main focus of our network targets improving the perceptual quality of reconstructed image. Thus, we designed a perceptual metric guided deep attention network (PM-DAN) for achieving this goal. The details of our network will be presented in the next section.

3. Perceptual Metric Guided Deep Attention Network

Figure 1 shows the proposed PM-DAN for a single image SR. Similar to DIP [11], we take the output $f_{\Theta}(z)$ of a parametric generator G to represent the unknown HR image $x^h \in R^{C \times H \times W}$, in which the random noise tensor $z \in R^{C' \times H \times W}$ is the network input and Θ is the network parameters. z has the same spatial resolution as the network output x^h . C is the channel number of x^h , and it is set as 3 for the color image. In the case of supervised learning, the network parameters are usually learned from the training set under the objective function that minimizes the mean reconstruction error. Unlike previous work, we optimize the network parameters according to the image resolution

degradation model to ensure that the output of the generator $x^h = f_{\Theta}(z)$ can match with the given LR image x^l , and the objective function of network learning is formulated as,

$$\begin{aligned} x &= \min_{\Theta} L^P(x^l - Df_{\Theta}(z)) \\ &= \min_{\Theta} \alpha L^{MAE}(x^l - Df_{\Theta}(z)) + (1 - \alpha) L^{MSSIM}(x^l - Df_{\Theta}(z)) \end{aligned} \quad (1)$$

where D is the down-sampling operator for image resolution degeneration, L^P is the perceptual metric consisting of the mean absolute error metric L^{MAE} and multi-scale structural similarity metric L^{MSSIM} [13], and $\alpha \in (0, 1)$ is the regularization weight. The weights Θ are learned to minimize the perceptual metric given a specific LR image x^l , thereby boosting visual quality of the reconstructed image.

3.1. Network Architecture

As shown in the bottom half of Figure 1, our generator network G has an attention-based encoder-decoder architecture consisting of three types of modules, namely a down-scale module, a skip connection module, and an up-scale module. The detailed configurations of our generator network G are shown in Table 1. The down-scale module in encoder is used to extract multi-scale features, the skip connection module delivers feature maps from encoder to decoder via convolution and concatenation operation, and the up-scale module in decoder is responsible for conducting reconstruction at different scales. Each convolution layer in these modules is coupled with batch normalization (BN) and nonlinear LeakyReLU (0.2) activation, and the kernel size of convolutional layers is set as 3×3 . Different from Reference [11], we enhance the up-scale module by inserting two residual spatial attention (RSA) units. Under the guidance of the perceptual metric, it is expected that the predicted spatial attention maps will highlight areas with rich visually sensitive structures. Therefore, the up-scale module can pay more attention to informative features at different scales for reconstruction.

Table 1. The configurations of our generator network G .

Block	Layer	Name	Parameters<Kernel-Inchannel-Outchannel-Padding-Stride>
Input	In-1	Conv	3-64-128-1-1
Down-scale module	Down-1	Conv	3-128-128-1-2
	Down-2	Conv	3-128-128-1-1
Skip module	Skip-1	Conv	3-128-64-1-1
	Skip-2	Conv	1-64-4-1-1
Up-scale module	Up-1	Conv	3-132-128-1-1
	RSA	Conv	3-128-128-1-1
		Conv	3-128-128-1-1
		DilatedConv	3-128-1-3-1 <dilation = 3>
	Up-2	Conv	1-128-128-0-1
Output	Out-1	Conv	1-128-3-0-1

The inner diagram of RSA unit is shown in Figure 2. Motivated by Reference [26,27], RSA adopts residual learning mechanism, and the output of RSA is computed as the sum of input and input masked by the predicted attention map. The mathematical formulation of RSA is

$$\begin{aligned} X_c &= W^2 * \delta(W^1 * F_{c-1}), \\ F_c &= F_{c-1} + f_c(X_c) \odot X_c, \end{aligned} \quad (2)$$

where $'*$ ' represents the convolution operation, F_{c-1} is the input of RSA, X_c is an intermediate result computed from F_{c-1} through the operation flow of convolution W^1 , ReLU activation function δ [28] and convolution W^2 , $f_c(\cdot)$ predicts the spatial attention map from X_c , $\odot$ is the point-wise multiplication, and F_c is the final output of RSA. The spatial attention map $f_c(X_c)$ is computed as,

$$\begin{aligned} f_c(X_c) &= \text{sigmoid}(W^d * X_c) \\ &= \frac{1}{1 + \exp(-W^d * X_c)}, \end{aligned} \quad (3)$$

where W^d is a 3×3 dilated convolution [29] with the dilation rate of 3, and $f_c(X_c)$ is the obtained single-channel attention map. By enlarging the receptive field through dilated convolution, a larger range of information can be used to predict response in the attention map. By using a residual link, cascading two RSA units does not cause attenuation of the response values in the feature map. In contrast, RSA units not only increase the depth of the network but also enable the network to focus on important features, thereby improving the quality of the reconstructed image.

Figure 2. The diagram of residual spatial attention (RSA) unit.

3.2. Loss Function

According to the study of Reference [13], we take the perceptual metric defined in Equation (1) as the loss layer to drive our attention-based network learning, thereby preserving the visually sensitive structures in the HR image. The first loss term L^{MAE} in L^P is l_1 norm, which sums the absolute error at each pixel p . The mathematical formulation is defined as:

$$L^{MAE}(x^l, y) = \frac{1}{N} \sum_{p=1}^N |x^l(p) - (Dx^h)(p)|, \quad (4)$$

where $x^l(p)$ is the pixel value of x^l at the position p , N is the total number of pixels in x^l , and Dx^h is the downsampled image from x^h and denotes as y . The second loss term in L^P exploits MSSSIM metric [12] to measure the reconstruction error between x^l and y . MSSSIM is a multi-scale generalization of SSIM metric. Before introducing the mathematical formula of MSSSIM, we first give the definition of SSIM metric as,

$$\begin{aligned} SSIM(x^l, y) &= \frac{2\mu_x\mu_y + C_1}{\mu_x^2\mu_y^2 + C_1} \times \frac{2\sigma_{xy} + C_2}{\sigma_x^2\sigma_y^2 + C_2} \\ &= l(x^l, y) \times cs(x^l, y). \end{aligned} \quad (5)$$

By iteratively filtering and downsampling of the input image by $M - 1$ times, we can obtain M scales of the input image, and, accordingly, MSSSIM calculates structural similarity by combining the measurement of M scales,

$$MSSSIM(x^l, y) = l_M(x^l, y) \times \prod_{j=1}^M cs_j(x^l, y). \quad (6)$$

Therefore, the loss L^{MSSSIM} is set to 1 minus the negative MSSSIM metric,

$$L^{MSSSIM}(x^l, y) = \frac{1}{N} \sum_P 1 - MSSSIM(P). \quad (7)$$

In Equations (5) and (6), l_j is the divergence in brightness, cs_j is the compound divergence in contrast and structure at scale $j = 1, \dots, M$, μ_x and σ_x represent the mean and standard deviations of the patch P centered at a pixel p of x^l , respectively, μ_y and σ_y correspond to the mean and standard deviations of y at the pixel p , respectively, σ_{xy} denotes the covariance of x^l and y , and C_1, C_2 are small positive constants which can avoid the case of dividing by zero. The mean and standard variance associated with the patch P are calculated by a convolution with Gaussian kernel G_σ with the standard variance σ . The subscript p is omitted in MSSSIM metric for simplicity. N is the total number of patches produced by sliding the patch along the whole image y .

In order to propagate the reconstruction error from the loss layer to the previous layers, we need to first define the derivative of L^P loss. Specifically, the derivative of L^{MAE} for back-propagation can be calculated as,

$$\frac{\partial L^{MAE}(p)}{\partial p} = D' \text{sign}(x^l(p) - y(p)), \quad (8)$$

where D' is the transpose of the downsampling matrix D . The calculation of MSSSIM for each patch P involves neighborhood pixels of the pixel p . According to the chain rule, we need to calculate the derivative of $L^{MSSSIM}(P)$ at the pixel p with respect to all the other pixels p' in the patch P , and the derivation formula is

$$\begin{aligned} \frac{\partial L^{MSSSIM}(p)}{\partial y(p')} &= D' \left[\sum_{p \in P} - \frac{\partial}{\partial y(p')} MSSSIM(p) \right] \\ &= D' \left[\sum_{p \in P} - \left(\frac{\partial l_M(p)}{\partial y(p')} + l(p) \times \sum_{i=1}^M \frac{1}{cs_i(p)} \frac{\partial cs_i(p)}{\partial y(p')} \right) \times \prod_{j=1}^M cs_j(p) \right], \end{aligned} \quad (9)$$

where $l(p)$ and $cs(p)$ are corresponding to the brightness divergence and compound divergence of contrast and structure at the pixel p , namely the first and second term of Equation (5). Their derivation details can be referred to the additional material in Reference [13].

The derivatives of the perceptual metric L^P can hence simply calculated as the weighed sum of the derivatives of L^{MAE} and L^{MSSSIM} according to Equations (8) and (9). Adam algorithm is then used to minimize L^P and the optimal network parameters can be found for reconstruction. Different from supervised learning over a given training set in Reference [13], our network is optimized for SR reconstruction from only a given LR observation.

4. Experimental Results and Analysis

We conduct experiments on the Set5 [30], Set14 [31] and two images from the Internet to validate the performance of the proposed PM-DAN. The height and width of the full-resolution images in these two datasets range from 228 to 768. First, we test the impact of the hyper-parameters (including the weight α in the perceptual metric and the iteration number in network learning) on the reconstruction results. Then, some ablation studies are conducted to verify whether

the attention-based network and perceptual metric are beneficial for SR reconstruction. Finally, PM-DAN is compared quantitatively and qualitatively with bicubic interpolation, DIP [11], SRCNN [15], and LapSRN [23]. PSNR and SSIM are used as quantitative metrics for measuring reconstruction quality. The source codes of DIP, SRCNN, and LapSRN are downloaded from online websites provided by the authors. The parameters of DIP, SRCNN, and LapSRN are set to be the same as the default values in the source code. The proposed PM-DAN is based upon the Pytorch [32] framework and runs on NVIDIA RTX 2080 GPU. The channel of input random noise tensor Z is set as 64. Adam is used for our network learning [33], and the learning rate is set to 0.001.

4.1. Parameters Analysis

The weight α in the perceptual metric. α is used to balance the importance of l^{MAE} loss and l^{MSSSIM} in the perceptual metric. Thus, we uniformly sample α at the interval of 0.05 in the range of 0 to 1. Figure 3 shows the curves of the mean PSNR and SSIM values versus different α upon three images from the Set14 [31] in the case of $4\times$ super resolution. When α is approximately equal to 0.16, the proposed method achieves the best PSNR and SSIM values. It implies that L^P is better than either l^{MAE} or l^{MSSSIM} for improving reconstruction quality, and it also verifies the rationality of the hybrid of l^{MAE} and l^{MSSSIM} . Thus, α is set as 0.16 in the subsequent experiments.

The iteration number in network learning. Both PM-DAN and DIP use iterative optimization to generate the HR images that match the LR observation as much as possible. The number of iterations will impact the final result. Figure 4 presents the PSNR and SSIM curves of PM-DAN and DIP versus iteration numbers upon the Zebra image from the Set14 in the case of $4\times$ super resolution. The maximum iteration number is set as 3000. It can be seen that both the PSNR and SSIM curves of PM-DAN and DIP increase rapidly before 1500 iterations, then rise slowly to 2000 iterations and saturate near 3000 iterations. Although the PSNR and SSIM curves of PM-DAN and DIP follow a similar trend, PM-DAN has a higher PSNR and SSIM than DIP. Taking account of the compromise between time complexity and reconstruction performance, we select 2000 iterations for PM-DAN and DIP in the following experiments.

Figure 3. The curves of PSNR and structural similarity (SSIM) values versus different α values. The point corresponding to the highest PSNR or SSIM value has been highlighted.

Figure 4. The curves of PSNR and SSIM values versus iteration numbers.

4.2. Ablation Studies

In this section, some ablation studies are performed to verify the strengthes of RSA units and the perceptual metric in the proposed PM-DAN. In detail, we implement another two simplified versions of PM-DAN, one without RSA units (PM-DAN w/o RSA) and one without the perceptual metric (PM-DAN w/o PL). We also compare PM-DAN and its two simplified versions with DIP. DIP can be regarded as a simplified PM-DAN model without RSA units and the perceptual metric. Table 2 presents the $4\times$ SR reconstruction results of these four algorithms upon the Set14. The best PSNR and SSIM values are highlighted in bold. We can see that two simplified versions of PM-DAN both have superior performance than DIP, which demonstrates the advantages of RSA units and the perceptual metric for improving SR quality. However, the deployment of the perceptual metric can only result in marginal improvement. PM-DAN has the best reconstruction results in terms of both PSNR and SSIM. This reveals that the joint deployment of RSA units and the perceptual metric can produce positive incentives and further improve reconstruction quality.

Table 2. The ablation studies to verify the influence of RSA unit and the perceptual metric on the PSNR and SSIM values of reconstruction images.

Image	DIP	PM-DAN w/o RSA	PM-DAN w/o PL	PM-DAN
	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM
Baboon	22.29/0.5195	22.59/0.5323	22.62/0.5419	22.68/0.5481
Barbara	25.53/0.7286	25.56/0.7310	25.60/0.7358	25.77/0.7472
Bridge	23.09/0.5861	23.45/0.5701	23.41/0.5617	23.68/0.5914
Coastguard	25.81/0.6490	26.00/0.6169	25.89/0.6351	26.05/0.6415
Comic	22.18/0.6889	22.43/0.6866	22.49/0.7016	22.58/0.7075
Face	31.02/0.7507	31.97/0.7927	32.01/0.7944	32.11/0.8014
Flowers	26.14/0.7617	26.63/0.7839	26.65/0.7880	26.93/0.7998
Foreman	31.66/0.8845	31.96/0.9010	31.84/0.8970	32.49/0.9082
Lenna	30.83/0.8367	31.15/0.8487	31.27/0.8498	31.36/0.8556
Man	26.09/0.7079	26.49/0.7280	26.57/0.7405	26.75/0.7507
Monarch	29.98/0.9083	29.77/0.9093	30.02/0.9159	30.39/0.9236
Pepper	32.08/0.8524	32.23/0.8599	32.45/0.8646	32.77/0.8708
Ppt3	24.38/0.8815	24.31/0.8832	24.74/0.8906	25.10/0.9050
Zebra	25.71/0.7477	26.02/0.7777	26.21/0.7791	26.53/0.7871
AVG.	26.91/0.7503	27.18/0.7588	27.27/0.7640	27.51/0.7742

Figure 5 shows the reconstructed HR images of Lenna and Man by PM-DAN and PM-DAN without the perceptual metric. As can be seen, when *mse* loss is used as the loss layer, the obtained SR image will become blurry, and many details are lost. Conversely, by utilizing the perceptual metric, the reconstructed SR images can have sharp edge and contour structures. The zooming-in visualization

of Lenna's hat and Man's face demonstrates the effectiveness of the perceptual metric for preserving image structures.

The SR images of Barbara and Comic by PM-DAN and PM-DAN without RSA units are shown in Figure 6. The multi-scale spatial attention maps predicted by RSA units are also presented. We can see that the attention maps at different scales exhibit high response intensity in different areas, and the union of these high-response areas can almost cover the entire image. With the progressive refinement of the scale of the attention map, the areas with high response intensity mainly concentrate on the flat and local structures of image, which are consistent with sensitive characteristics of HVS. Due to contrast masking phenomenon of HVS [34], reconstruction distortions in structural areas are more likely to be perceived than texture regions. With the aid of RSA units, PM-DAN can well localize the visually sensitive areas at different scales, thus the visually informative structures can be preserved in the reconstructed image, especially in the area with highlighted attention response. This also explains why the combination of the perceptual metric and attention units can produce the better reconstruction results. Taking the image comic as an example, the spatial attention map at the finest scale has high response strength in the area of Girl's chin. Accordingly, Girl's chin is reconstructed with enhanced visual quality.

Figure 5. The ablation studies of perceptual metric for reconstruction. The images from left to right correspond to the ground truth, PM-DAN, and PM-DAN without the perceptual metric.

Figure 6. The ablation studies of RSA unit for reconstruction. The images from left to right correspond to ground truth, PM-DAN without RSA units, PM-DAN, and the predicted spatial attention maps by RSA units at multiple scales.

4.3. Performance Comparison

We compare the proposed PM-DAN with bicubic interpolation, DIP [11], SRCNN [15], and LapSRN [23]. DIP and PM-DAN do not require an image set to pre-train the models, while SRCNN uses a large training set consisting of 395,909 images from the ILSVRC2013 ImageNet detection training partition, and LapSRN employs 91 images from [7] and 200 images from BSD200 [35] as the training data for learning the reconstruction mapping. The symbols T and NT are used to represent the methods with or without pre-training, respectively. Tables 3–6 show quantitative PSNR and SSIM values of multiple methods for $4\times$ and $8\times$ SR upon the Set5 and Set14. The best PSNR and SSIM values are highlighted in bold.

Table 3. $4\times$ SR comparison on Set5. The best PSNR and SSIM values are highlighted in bold.

Set5 $\times 4$	Bicubic(NT) PSNR/SSIM	DIP(NT) PSNR/SSIM	PM-DAN(NT) PSNR/SSIM	SRCNN(T) PSNR/SSIM	LapSRN(T) PSNR/SSIM
Baby	31.78/0.8365	31.49/0.8589	32.65/0.8881	33.13/0.8835	33.55/0.9044
Brid	30.20/0.8496	31.80/0.9052	32.83/ 0.9265	32.52/0.9095	33.76/0.9063
Butterfly	22.13/0.7542	26.23/0.8805	26.32/0.8811	25.44/0.8503	27.28/0.8883
Head	31.34/0.7820	31.04/0.7609	31.97/0.7962	32.45/0.7817	32.62/0.8101
Woman	26.75/0.8299	28.93/0.8788	29.47/0.9021	28.88/0.8542	30.72/0.9159
AVG.	28.44/0.8104	29.89/0.8568	30.65/0.8788	30.48/0.8558	31.59/0.8850

Table 4. $8\times$ SR comparison on Set5. The best PSNR and SSIM values are highlighted in bold.

Set5 $\times 8$	Bicubic(NT) PSNR/SSIM	DIP(NT) PSNR/SSIM	PM-DAN(NT) PSNR/SSIM	LapSRN(T) PSNR/SSIM
Baby	27.28/0.7166	28.28/0.7548	28.84/0.7645	28.88/0.7701
Brid	25.28/0.7015	27.09/ 0.7628	26.92/0.7580	27.10/0.7615
Butterfly	17.74/0.5661	20.02/0.6705	20.60/0.6811	19.97/0.6789
Head	28.82/0.6016	29.55/0.6879	29.52/0.6941	29.76/0.7103
Woman	22.74/0.7043	24.50/0.7555	24.77/0.7635	24.79/0.7692
AVG.	24.37/0.6580	25.88/0.7263	26.13/0.7322	26.10/0.7380

Table 5. $4\times$ SR comparison on Set14. The best PSNR and SSIM values are highlighted in bold.

Set14 $\times 4$	Bicubic(NT) PSNR/SSIM	DIP(NT) PSNR/SSIM	PM-DAN(NT) PSNR/SSIM	SRCNN(T) PSNR/SSIM	LapSRN(T) PSNR/SSIM
Baboon	22.44/0.4712	22.29/0.5195	22.68/ 0.5481	22.72/0.5015	22.83 /0.5372
Barbara	25.15/0.6793	25.53/0.7286	25.77 /0.7472	25.75/0.7322	25.69/ 0.7454
Bridge	22.96/0.5328	23.09/0.5861	23.68/0.5914	23.75 /0.5955	23.74/ 0.6203
Coastguard	25.53/0.5353	25.81/0.6490	26.05/ 0.6415	26.03/0.5610	26.21 /0.6016
Comic	21.59/0.5650	22.18/0.6889	22.58/ 0.7075	22.69/0.6701	22.90 /0.7067
Face	31.34/0.7440	31.02/0.7507	32.11/ 0.8014	32.37/0.7796	32.62 /0.7996
Flowers	25.33/0.7126	26.14/0.7617	26.93/ 0.7998	27.13/0.7821	27.54 /0.7925
Foreman	29.45/0.8654	31.66/0.8845	32.49/0.9085	32.11/0.8991	33.59 / 0.9219
Lenna	29.84/0.8139	30.83/0.8367	31.36/ 0.8556	31.40/0.8453	31.98 /0.8543
Man	25.70/0.6677	26.09/0.7079	26.75/0.7507	26.88/0.7303	27.27 / 0.7624
Monarch	27.45/0.8923	29.98/0.9083	30.39/ 0.9236	30.21/0.9193	31.62 /0.9230
Pepper	30.63/0.8427	32.08/0.8524	32.77/ 0.8708	32.97/0.8673	33.88 /0.8551
Ppt3	21.78/0.8353	24.38/0.8815	25.10/0.9045	24.79/0.8964	25.36 / 0.9119
Zebra	24.01/0.6799	25.71/0.7477	26.53/ 0.7871	26.08/0.7488	26.98 /0.7758
AVG.	25.92/0.7027	26.91/0.7503	27.51/ 0.7742	27.49/0.7520	27.97 /0.7720

Table 6. $8\times$ SR comparison on Set14. The best PSNR and SSIM values are highlighted in bold.

Set14 $\times 8$	Bicubic(NT) PSNR/SSIM	DIP(NT) PSNR/SSIM	PM-DAN(NT) PSNR/SSIM	LapSRN(T) PSNR/SSIM
Baboon	21.28/0.3292	21.37/0.3688	21.46/0.3694	21.51 / 0.3744
Barbara	23.44/0.5649	23.90/0.6153	24.04/0.6168	24.21 / 0.6231
Bridge	21.54/0.3614	21.58/0.3970	22.13 /0.4001	22.11/ 0.4097
Coastguard	23.65/0.4028	24.17/0.4236	24.32 /0.4300	24.10/ 0.4303
Comic	19.25/0.3848	19.79/0.4498	20.04/0.4531	20.06 / 0.4579
Face	28.79/0.6589	29.48/0.6915	29.58/0.6945	29.85 / 0.7092
Flowers	22.06/0.5539	22.93/0.5953	22.93/ 0.5960	23.31 /0.5941
Foreman	25.37/0.7587	27.01/0.8223	28.16 / 0.8224	28.13/0.8217
Lenna	26.27/0.7053	27.72/0.7553	28.00/0.7572	28.22 / 0.7637
Man	23.06/0.5247	23.92/0.5639	23.88/0.5724	24.20 / 0.5789
Monarch	23.18/0.7753	24.02/0.8085	24.98 /0.8093	24.97/ 0.8147
Pepper	26.55/0.7406	28.63/0.7975	29.01/0.7980	29.22 / 0.8058
Ppt3	18.62/0.7062	20.09/0.7606	20.52 /0.7671	20.13/ 0.7717
Zebra	19.59/0.4572	20.25/0.5086	21.05 /0.5241	20.28/ 0.5253
AVG.	23.04/0.5660	23.91/0.6112	24.27/0.6150	24.31 / 0.6200

In the four cases of experiments, the PSNR and SSIM values of the proposed PM-DAN are all better than DIP. Moreover, PM-DAN has better PSNR and SSIM values than the pre-trained SRCNN in the case of $4\times$ SR upon Set5 and Set14. PM-DAN can also achieve comparable results with the pre-trained LapSRN, and even outperforms LapSRN in some cases, such as the averaged PSNR value in the case of $8\times$ SR upon the Set5 and the averaged SSIM value in the case of $4\times$ SR upon the Set14. Some $4\times$ and $8\times$ reconstructed images are shown in Figures 7 and 8, respectively. The zooming-in display of different patches are also presented. With the aid of pre-training, the SR results of LapSRN also show good visual quality. PM-DAN can reconstruct the HR image with sharp structures and texture details, and has better visual quality than DIP. The brightness and color of the reconstructed image can also be well preserved by PM-DAN. In the case of $8\times$ SR, PM-DAN can even recover the clearer structures than LapSRN, such as the eyes in the Baby image and the spot textures in the Butterfly image.

We further evaluate the performance of our method by conducting the SR experiments on two real images from the internet, one remote sensing image and one landscape image. Figure 9 presents the $4\times$ SR results of our method and DIP upon these two images. The resolutions of the $4\times$ SR images of these two images are 864×576 and 1088×736 , respectively. We can see that, compared to the DIP, our method can recover sharp structures and more texture details. As shown in the zooming-in patches of the remote sensing image, our method can reconstruct more details in the areas of house and wood. With regard to the landscape image, the reconstructed image of our method has good contrast and saturation, which makes the reconstructed image visually attractive.

Figure 7. The visualization of $4\times$ super resolution results.

Figure 8. The visualization of $8\times$ super resolution results.

Figure 9. The visualization of $4\times$ super resolution results of our method and Deep Image Prior (DIP) on two real images.

5. Conclusions

In this paper, we proposed a unsupervised SR network named PM-DAN. An attention-based decoder-encoder network is designed to predict the SR reconstruction, in which residual spatial attention units are deployed in each decoding layer to concentrate informative feature for reconstruction. Meanwhile, the network is learned under the guidance of the perceptual metric, which has good potential of recovering visually sensitive structures. The experimental results demonstrate that PM-DAN effectively improves the visual quality of SR image and can outperform DIP in terms of both PSNR and SSIM, even producing comparable results with the pre-trained LapSRN network. In future work, we plan to combine our model with appropriate domain-specific regularization to obtain better SR results.

Author Contributions: Conceptualization, Y.S. (Yubao Sun) and W.Z.; Supervision, W.Z.; software, Y.S. (Yubao Sun) and Y.Y. (Yuyang Shi); writing—original draft preparation, Y.S. (Yubao Sun) and Y.S. (Yuyang Shi); writing—review and editing, Y.S. (Yubao Sun) and Y.S. (Yuyang Shi); Data curation, Y.S. (Yuyang Shi) and Y.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant numbers (61672292), in part by the Key University Science Research Project of Jiangsu Province under Grant Number 18KJA520007, in part by Six Talent Climax Foundation of Jiangsu under Grant Number 2016-DZXX-037.

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

References

1. Pouliot, D.; Latifovic, R.; Pasher, J.; Duffe, J. Landsat Super-Resolution Enhancement Using Convolution Neural Networks and Sentinel-2 for Training. *Remote Sens.* **2018**, *10*, 394. [\[CrossRef\]](#)
2. Cherukuri, V.; Guo, T.; Schiff, S.; Monga, V. Deep MR Brain Image Super-Resolution Using Spatio-Structural Priors. *IEEE Trans. Image Process.* **2020**, *29*, 1368–1383. [\[CrossRef\]](#) [\[PubMed\]](#)
3. Keys, R. Cubic convolution interpolation for digital image processing. *IEEE Trans. Acoust. Speech Signal Process.* **1981**, *29*, 1153–1160. [\[CrossRef\]](#)
4. Sajjadi, M.S.M.; Scholkopf, B.; Hirsch, M. Enhancenet: Single image super-resolution through automated texture synthesis. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 4491–4500.
5. Kim, K.I.; Kwon, Y. Single-image super-resolution using sparse regression and natural image prior. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *32*, 1127–1133. [\[PubMed\]](#)
6. Freeman, W.T.; Jones, T.R.; Pasztor, E.C. Example-based super-resolution. *IEEE Comput. Graph. Appl.* **2002**, *22*, 56–65. [\[CrossRef\]](#)
7. Yang, J.; Wright, J.; Huang, T.; Ma, Y. Image superresolution via sparse representation. *IEEE Trans. Image Process.* **2010**, *19*, 2861–2873. [\[CrossRef\]](#) [\[PubMed\]](#)
8. Sun, Y.; Chen, J.; Liu, Q.; Liu, G. Learning image compressed sensing with sub-pixel convolutional generative adversarial network. *Pattern Recognit.* **2020**, *98*, 107051. [\[CrossRef\]](#)
9. Li, K.; Wu, Z.; Peng, K.C.; Ernst, J.; Fu, Y. Tell me where to look: Guided attention inference network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 9215–9223.
10. Yang, W.; Zhang, X.; Tian, Y.; Wang, W.; Xue, J.H.; Liao, Q. Deep Learning for Single Image Super-Resolution: A Brief Review. *IEEE Trans. Multimed.* **2019**, *21*, 3106–3121. [\[CrossRef\]](#)
11. Ulyanov, D.; Vedaldi, A.; Lempitsky, V. Deep image prior. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 9446–9454.
12. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [\[CrossRef\]](#) [\[PubMed\]](#)
13. Zhao, H.; Gallo, O.; Frosio, I.; Kautz, J. Loss functions for neural networks for image processing. *IEEE Trans. Comput. Imaging* **2017**, *3*, 47–57. [\[CrossRef\]](#)
14. Viet, K.H.; Ren, J.; Xu, X.; Zhao, S.; Xie, G.; Vargas, V.M. Deep Learning Based Single Image Super-resolution: A Survey. *Int. J. Autom. Comput.* **2019**, *16*, 413–426.

15. Dong, C.; Loy, C.C.; He, K.; Tang, X. Image super-resolution using deep convolutional networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *38*, 295–307. [[CrossRef](#)] [[PubMed](#)]
16. Kim, J.; Lee, J.K.; Lee, K.M. Accurate image super-resolution using very deep convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1646–1654.
17. Zhang, K.; Zuo, W.; Gu, S.; Zhang, L. Learning deep CNN denoiser prior for image restoration. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3929–3938.
18. Kim, J.; Lee, J.K.; Lee, K.M. 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.
19. Dong, C.; Loy, C.C.; Tang, X. Accelerating the Super-Resolution Convolutional Neural Network. *Eur. Conf. Comput. Vis.* **2016**, 391–407. [[CrossRef](#)]
20. Shi, W.; Caballero, J.; 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 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 IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 July 2017; pp. 136–144.
22. 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 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4681–4690.
23. 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 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 624–632.
24. Feng, X.; Su, X.; Shen, J.; Jin, H. Single Space Object Image Denoising and Super-Resolution Reconstructing Using Deep Convolutional Networks. *Remote Sens.* **2019**, *11*, 1910. [[CrossRef](#)]
25. Zhang, H.; Sindagi, V.; Patel, V.M. Image de-raining using a conditional generative adversarial network. *IEEE Trans. Circuits Syst. Video Technol.* **2019**. [[CrossRef](#)]
26. 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.
27. Wang, F.; Jiang, M.; Qian, C.; Yang, S.; Li, C.; Zhang, H.; Wang, X.; Tang, X. Residual attention network for image classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3156–3164.
28. Glorot, X.; Bordes, A.; Bengio, Y. Deep sparse rectifier neural networks. In Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, Ft. Lauderdale, FL, USA, 11–13 April 2011; pp. 315–323.
29. Yu, F.; Koltun, V. Multi-scale context aggregation by dilated convolutions. *arXiv* **2015**, arXiv:1511.07122.
30. Bevilacqua, M.; Roumy, A.; Guillemot, C.; Marie, A.M. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In Proceedings of the 23rd British Machine Vision Conference, Surrey, UK, 3–7 September 2012.
31. Zeyde, R.; Elad, M.; Protter, M. On single image scale-up using sparse-representations. *Int. Conf. Curves Surfaces* **2010**, 711–730. [[CrossRef](#)]
32. 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 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017.
33. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.

34. Legge, G.E.; Foley, J. Contour detection and hierarchical image segmentation. *J. Opt. Soc. Am.* **1980**, *70*, 1458–1471. [[CrossRef](#)] [[PubMed](#)]
35. Arbelaez, P.; Maire, M.; Fowlkes, C.; Malik, J. Contour detection and hierarchical image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *33*, 898–916. [[CrossRef](#)] [[PubMed](#)]

© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).