Superpixel Driven Unsupervised Deep Image Super-resolution

Jun Yang (yangjun1@stu.xhu.edu.cn)
Xihua University
Chao Zhang
Sichuan Police College
Li Xu
Xihua University
Bing Luo
Xihua University

Research Article

Keywords: image super-resolution, Deep Image Prior, unsupervised, segmentation driven, entropy

Posted Date: October 28th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-2204044/v1

Superpixel Driven Unsupervised Deep Image Super-resolution

Jun Yang[1*], Chao Zhang[2], Li Xu[3] and Bing Luo[1]

[1]School of Computer and Software Engineering, Xihua University, Chengdu, 610039, China.
[2]key laboratory of intelligent policing, Sichuan Police College, Luzhou, 646000, China.
[3]School of Science, Xihua University, Chengdu, 610039, China.

*Corresponding author(s). E-mail(s): yangjun1@stu.xhu.edu.cn;
Contributing authors: galoiszhang@gmail.com;
xuli19860715@163.com; Mathild1987@163.com;

Abstract

Most of the existing deep learning-based image super-resolution methods require a large number of datasets or ground truth. However, these methods are not suitable for the restoration of real image with different domains. Recently, Deep Image Prior (DIP) based on single-image explores image prior and uses network structure as implicit image prior to recover images, but it ignores the explicit prior information of the actual image itself. The addition of image prior can effectively alleviate the ill-posed problem in the image restoration model. Therefore, in this paper, we propose an unsupervised deep image super-resolution (SR) method that based on segmentation driven. Intuitively, clear image has a clearer segmentation boundary. It will drive deep neural networks (DNN) to obtain higher performance SR image when forcing the restored image to have clear boundary. In order to make energy flow into DIP better, we use the fully convolutional networks-based (FCN-based) superpixel method, and we use back propagation to inject the gradient generated by segmentation entropy energy into DIP to obtain lower energy optimization parameters. Experiments show that the image generated by our method has clearer boundary and better performance than that generated by DIP on Set5, Set14 and BSD100.

Keywords: image super-resolution; Deep Image Prior; unsupervised; segmentation driven; entropy

1 Introduction

Image super-resolution reconstruction technology is to generate a high-resolution image from low-resolution images by algorithm. However, the resolution of the image will be limited by signal transmission bandwidth and noise interference. Restricted by the current manufacturing level, it is difficult to improve the resolution from the hardware. So people pay more and more attention to the research of image super-resolution algorithm. Now, super-resolution algorithm is widely used in medical imaging [1–3], remote sensing imaging [4–6] and picture compression [7–9].

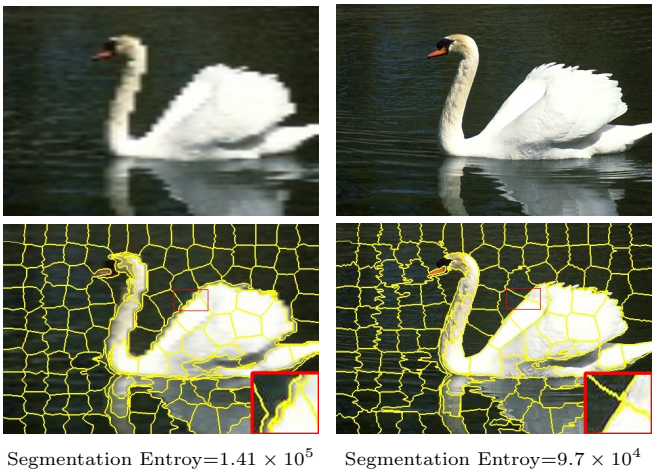

Fig. 1 The first line shows the images with blurred boundary and clear boundary, respectively. The second line shows the corresponding superpixel segmentation results. Based on FCN-based superpixel method, the segmentation is realized by predicting the probability of which superpixel pixel belongs to. The clearer the image is, the greater the probability of dividing the correct superpixel block is, and the clearer the segmentation boundary is. We convert this probability into a segmentation entropy, as shown in the third line. The clearer the image is, the smaller the obtained segmentation entropy is.

The existing methods to solve the problem of image super resolution are mainly divided into three parts. Methods based on interpolation, reconstruction and deep learning. The interpolation method [10–12] mainly use the relationship between adjacent pixels of the image to select the appropriate pixel coordinates for image interpolation. However, these methods do not consider the semantic information of the whole image. Because of only using the value between the adjacent pixels of the original image to improve the resolution, the edge of the reconstructed image is poor. The method based on reconstruction mainly introduces effective prior information in the reconstruction process and uses prior information to constrain the image. Such as [13], it references probability priors to image super-resolution, [14] applies image gradient contour information to image edge contour restoration. The method

based on reconstruction obtain less prior knowledge when the up-sampling factor of the image is too large, so the reconstruction effect will decline sharply. The method based on deep learning is to mine the detail features of the image by convolutional neural network, it can recover the image by it's strong nonlinear fitting ability. Supervised SR methods [15–19] use paired low-resolution images and high-resolution images to learn the mapping from low-resolution images to high-resolution images. However, these methods require a large number of datasets and ground truth (GT). Unsupervised methods do not use GT, they are trained by some features of the image itself. such as [20–23] learn the degradation process by learning the domain between different images. [24, 25] train the super-resolution network model based on frequency separation. [26, 27] takes the network structure as the prior condition. [28, 29] use the image statistics within a single image. Unsupervised methods are learning the image degradation process to make the restored image more in line with the domain distribution of real images. However, these methods ignore the processing of image boundary details. Superpixels can get multiple superpixel blocks and edges by dividing pixels. It also can extract more segmentation boundary information.

In this paper, we propose an unsupervised deep image SR method that based on segmentation driven. Intuitively, clear image has a clearer segmentation boundary. We use superpixel segmentation method to extract the contour boundary of image and force them to become clear. The gradient generated by the segmentation entropy energy will flow into the neural network of DIP [26] through back propagation when forcing DIP to obtain a higher performance SR image. Specifically, we add segmentation entropy to each iteration of DIP as a driver, it can make DIP to focus more on edge detail recovery in each iteration and a clear edge contour image will be obtained. Experiments on Set5, Set14 and BSD100 datasets show that the image generated by our method has clearer boundary and better performance than that generated by DIP.

The main contributions of this paper are summarized as follows.

• We propose an unsupervised deep image SR method that based on segmentation driven and use back propagation to inject the gradient generated by segmentation entropy energy into DIP.

• In deep learning, a method for computing segmentation entropy is proposed.

• Solving the problem of insufficient edge detail recovery in unsupervised super resolution method.

2 Related Works

The existing methods to solve the problem of image super resolution are mainly divided into three parts. Methods based on interpolation, reconstruction and deep learning.

The method based on interpolation mainly uses the relationship between adjacent pixels to select the appropriate pixel coordinates for image interpolation. Such as [10], the value of the interpolation point is the value of the pixel with the shortest euclidean distance from the interpolation point. However, the results obtained by this method have obvious sawtooth phenomenon and the amplification effect is not ideal. In order to solve it, [12] mainly implements the linear interpolation of four adjacent pixels from the vertical and horizontal directions to realize the image interpolation. The enlarged image sawtooth phenomenon is improved, but the edge is blurred. Then Li [11] proposed an interpolation method based on edge guidance, it assumed that low-resolution images and high-resolution images had the same edge information at the edge. The prediction coefficient of the optimal linear super-resolution mapping was derived by calculating the local covariance of the edge of the low-resolution image. This method solves the problem of image edge sharpening, but the algorithm complexity is high. The method based on interpolation do not consider the semantic information of the whole image, the reconstruction effect is limited.

The method based on reconstruction mainly uses prior information to constrain image restoration. Schultz [13] introduced maximum a posteriori probability estimation model based on probability theory into image super-resolution reconstruction. However, the obtained high-resolution image edge contour is smooth. To solve this problem, Sun [14] proposed an image prior reconstruction method based on edge guidance. This image prior statistics the gradient contour information of the image, it can effectively sharpen the image edge. The reconstruction effect depends on the consistency of the statistical model and the gradient contour of the image. But once the magnification of this reconstruction-based method is too large, the effect of reconstruction will fall sharply.

The method based on deep learning is to extract the deep features of the image by convolutional neural network, and recover the image by it's strong nonlinear fitting ability. It can be divided into supervised SR and unsupervised SR.

The supervised super-resolution method uses a large number of datasets and GT to learn the mapping relationship between low-resolution images and high-resolution images. Then predict high resolution images based on the learned mapping relationship. Dong [15] introduced the convolutional neural network into the field of image super-resolution and proposed the SRCNN network structure. Specifically, it uses an interpolation method to resize the image, and then the high resolution image is obtained by nonlinear mapping with three layers convolution network. The mapping relationship between low-resolution images and high-resolution images is learned by convolution neural network. However, The method of changing the size of the image by interpolation and sending it to the neural network for recovery has affected the performance of image restoration. Then sub-pixel convolution layer is proposed by [17] , it does not require an up-sampling process for a given low-resolution image,

but indirectly realizes the image amplification process through the sub-pixel convolution layer. It improves the reconstruction effect. However, the above methods are using mean square error as the target loss function, which will cause the image to be too smooth and lack of sensory image realism. So Ledig [16] proposed SRGAN and applied GAN [30] to super-resolution tasks. The high-level feature mapping of the VGG [31] network is used to define the new perceptual loss. This loss uses the discriminant to make the generated high-resolution image as visually similar to the ground truth. Recently, to exploit feature correlation for improved performance, channel attention and second-order channel attention are further introduced by RCAN [18] and SAN [19]. But the datasets used by these methods are obtained through known degradation processes. If the model trained on this dataset is applied to low-resolution images with different domains in the real world, the effect is often not good. So people pay more and more attention to unsupervised super-resolution methods.

Unsupervised SR realizes image restoration by learning the image degradation process. Such as [20], it proposes a two-stage process to learn the degradation process. Firstly, using unmatched LR-HR images to train a HR-to-LR GAN network to learn the degradation process, this is to obtain natural LR images from HR images to simulate real low resolution data. And then the LR-to-HR GAN network is trained using paired LR-HR images on the basis of the first GAN. But this method does not take into account the generated resolution image feature distribution. So [21] proposed a Cycle-in-Cycle (CinCGAN) structure, it let LR space and HR space as two domains. Using Cycle-in-Cycle structure to learn the mapping between each other. Firstly, the network maps the input images with noise and blur to a low-resolution space that conforms to the real-world feature distribution and has no noise. Then, the feature distribution of the output high-resolution image is compared with that of the mismatched high-resolution image in the real world. This not only takes into account the feature distribution of low-resolution images, but also the feature distribution of high-resolution images. But this part of the SR model is a pre-trained model. In order to solve this problem, Maeda [22] starts from the high-resolution image, down-samples the high-resolution image, maps the down-sampled image to the real low-resolution domain, then passes through an up-sampling network to obtain a high-resolution image. [22] also compares the difference between the obtained high-resolution image and the real image in the domain. Recently, in order to better reduce domain bias, Wei [23] proposed domain distance map, which should be given different importance based on the distance from different regions to the target domain.

FSSR [24] and Zhou et al [25] proposed to learn a downsampling process to generate paired data and train SR network with the generated data in a supervised manner. FSSR [24] proposes frequency separation, which guides the network to realize the domain migration of high-frequency components, and uses the migrated images for SR network training. Zhou et al [25] is improved on the basis of FSSR [24]. A color-guided domain mapping networkwas proposed to alleviate the color shift in domain transformation process. Moreover,

it modified the discriminator of the super-resolution stage so that the network not only keeps the high frequency features, but also maintains the low frequency features.

However, these methods do not use the prior information of the image to constrain. Then DIP [26] uses the randomly initialized CNN as a prior. The deep image prior is to take into account that the CNN structure is sufficient to capture a large number of low-level image statistical priors. It takes random vector z as input and tries to generate target HR images. Since the network is randomly initialized and never trained, the only prior is the CNN structure itself. However, this method takes the network as an implicit prior information, and the restored image effect is not very good. Then, EIP [27] improves the DIP by introducing an external high-resolution reference image to enhance the image prior and update the input noise.

After that, the method of learning the degradation process based on the internal information of the image takes into account that the image statistics within a single image have provided sufficient information for SR. ZSSR [28] uses the nonlocal self-similarity of the image to exploit the internal recurrence of information within a single image. In the super-resolution of a low-resolution image, the image is down-sampled again to learn the super-resolution parameters between the LR image and the down-sampled LR image. Then it uses the parameters for LR super-resolution, we can get HR image finally. The method is based on internal learning and the mapping is learned from this image. The training time of ZSSR [28] is too long, and it needs thousands of iterations. To solve this problem, MZSR [29] present a novel training scheme based on meta-transfer learning, which learns an effective initial weight for fast adaptation to new tasks.

3 Methodology

In super-resolution tasks, although DIP can obtain high resolution images by training the network structure as an implicit prior, the effect is not ideal. Therefore, the idea of adding explicit prior to DIP is proposed to alleviate the ill-posed problem in image restoration model.

We observe that the clear image has a clearer segmentation boundary, it will driven DNN to obtain higher performance SR image when forcing the restored image to have clear boundary. The FCN-based superpixel method obtains multiple superpixel blocks and superpixel block edges by dividing pixels. In the FCN-based superixel method, some segmentation details are lost, because the segmentation loss is obtained by weighted averaging the predicted correlation matrix q and then reconstructing it. So we directly use the correlation matrix q and make the correlation matrix q into the segmentation entropy.

In order to obtain lower energy optimization parameters,we use back propagation to inject the gradient generated by segmentation entropy energy into DIP. Based on this theory, we calculate the segmentation entropy by convolution network, then add the segmentation entropy to each iteration of DIP,

it can force DNN to obtain a higher performance SR image. Fig. 1 shows the segmentation entropy results of image segmentation with clear edge and blurred edge. The results show that the clearer the boundary is, the lower the segmentation entropy is. To verify this idea. We tested on the BSD500 dataset. BSD500 is an image segmentation dataset which contains 500 complex natural images. As shown in Fig. 2, the dataset is processed with different degrees of ambiguity, and we calculate the segmentation entropy of images with different fuzzy degrees. The clearer the boundary, the smaller the segmentation entropy. We use different sampling methods (nearest, bilinear, bicubic) to downsample and then upsample the dataset. The image edge obtained by bicubic is clearer than that obtained by bilinear, so the segmentation entropy of the whole dataset is smaller. Since the image obtained by nearest has a sawtooth effect, the more obvious the sawtooth effect, the smaller the segmentation entropy. We downsample the dataset 2 times, 4 times, 6 times, 8 times, and then upsample to the original image size. Up_factor, Down_factor represents the operation of upsampling and downsampling the dataset according to different multiples. The abscissa represents the multiple of the downsampling and upsampling operations on the dataset. For bilinear and bicubic interpolations, the larger the value, the more blurred the image boundary we finally get. Vertical axis represents segmentation entropy result of dataset. We use convolutional neural network to calculate the correlation matrix q, and convert the correlation matrix q into the segmentation entropy. In order to prevent DIP from falling into local minimum, L2 regularization is added for constraint.

In this paper, we propose an unsupervised deep image SR method that based on segmentation driven. The reconstruction of super-resolution images can be divided into three steps. Firstly, random coding vector z is used as the input of DIP network, through the DIP network, we can get high resolution images as output. Secondly, the high resolution images is sent to the segmentation network for training to obtain the segmentation entropy. Finally, we use back propagation to inject the gradient generated by segmentation entropy energy into DIP, and the L2 regularization term is added for constraint. This method will be described in detail step by step in the following.

3.1 Deep Image Prior Network

The image is generated by $x = f_\theta(z)$. It maps the random code vector z to the image x and samples the real image from the random distribution. We interpret the neural network as a parametrization : $x = f_\theta(z)$, z is random coding vector, θ is the network parameter, x is the output result after parameter θ. In order to show the effect of parameterization, we consider the image inverse tasks, it can be expressed as the energy minimization problem, which is shown in Eq (1).

$$x^* = \min_x E(x;\ x_0) + R(x) \tag{1}$$

$E(x;\ x_0)$ is a task-dependent data term, $R(x)$ is the regularization term, which can usually capture the general prior of natural images, x^* is the image we want to get, x_0 is the image to be repaired. In this paper, the regularization

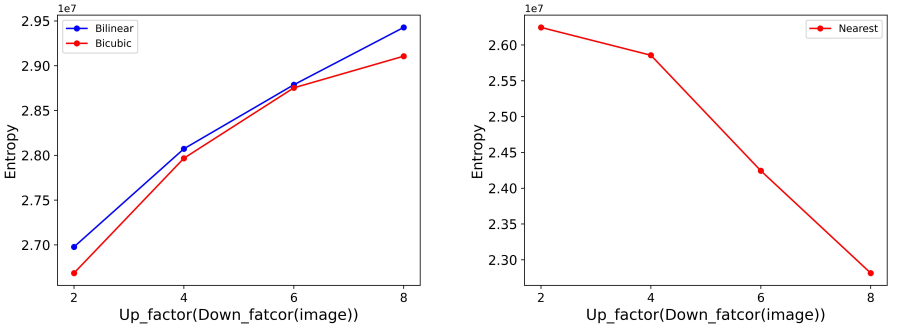

Fig. 2 In order to verify the effect of clear edge and fuzzy edge on segmentation entropy, we test on the BSD500 dataset. The results show that the clearer the segmentation boundary is, the smaller the entropy is.

term $R(x)$ is replaced by the implicit prior captured by the neural network, it lets the network learn the mapping from random vector z to degraded images. So, we can replace Eq (2) with Eq (1).

$$\theta^* = \underset{\theta}{\operatorname{argmin}} E\left(f_\theta(z);\ x_0\right), x^* = f_{\theta^*}(z) \tag{2}$$

Starting with the random initialization of parameters θ, then, in the next training, an optimizer such as gradient descent is used to obtain the minimum θ^*, the learned parameter θ^* is directly used to get the image x^*. z is a 2D tensor with 32 feature maps, the same size as x^*, and is filled with uniform noise. In Eq (2), $R(x)$ does not disappear, it is hidden, and it's value becomes an extreme form: $R(x) = 0$. The image can be speculated from z based on a specific CNN structure.

A high-capacity network can be used as the prior information, we expect to find a parameters θ that can reproduce the given target image x_0, including random noise. So that, the network should not impose any restrictions on the generated images. It's optimization is as Eq (3). The Eq (3) calculates the L_2-norm between x and x_0

$$E\left(x;\ x_0\right) = \left\|x - x_0\right\|^2 \tag{3}$$

Putting Eq (3) into Eq (2), it becomes an optimization problem Eq (4). For super resolution tasks, the data term is set as shown in Eq (5). The training process is shown in Algorithm 1.

$$\min_\theta \left\|f_\theta(z) - x_0\right\|^2 \tag{4}$$

$$E\left(x;\ x_0\right) = \min_\theta \left\|d(f_\theta(z)) - x_0\right\|^2, x = f_\theta(z) \tag{5}$$

$x_0 \subseteq R^{3\times H\times W}$ is a low resolution image, here H,W represent the height and width of the low-resolution image, respectively. Random code vector z

Algorithm 1 Algorithm for solving DIP

Input: Random code vector z and low-resolution image x_0.
Output: High-resolution image x.

repeat
  $x^{(i)} = \theta_x{}^{(i)}(z)$
  Update DIP_{loss} by Eq (5).
  Update $\theta_x{}^{(i)}$ using the ADAM algorithm.
  $i \leftarrow i + 1$
until $i > i_{max}$

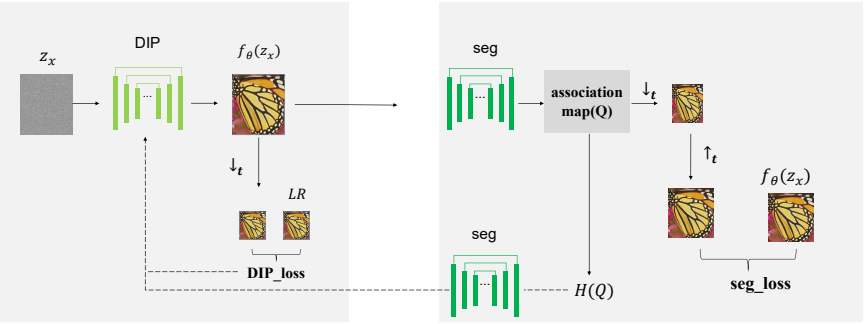

Fig. 3 The schematic illustration of training. Firstly, z_x is a random coding vector, the network structure is an encoder-decoder structure. f_θ represents the neural network, and the random coding vector is mapped to a high-resolution image through the neural network. $\downarrow_t$ denote downsampling with scale factor t. The image is reconstructed by minimizing the LR image, so as to optimize the model. Then, let HR image as the input, the segmentation network will output a correlation matrix q through an encoder-decoder neural network structure, which is a weight matrix to predict pixels to the surrounding nine clustering centers, here $\mathrm{q} \in \mathrm{Z}^{H \times W \times N_p}$. $\downarrow_t$ represents the process from the pixel point to the cluster center, implemented by Eq (6). t represents the grid pixel size of the clustering center as t. $\uparrow_t$ represents the process from superpixel center to pixel point, this is the process of image reconstruction, which is realized by Eq (7). Finally, minimize the input image and the reconstructed image. The training process is unsupervised. We emphasize that the two segmented networks in the figure are the same network. The gradient generated by the segmentation entropy energy flows into the DIP network through back propagation, and finally forces DNN to obtain better training parameters.

is the input of deep neural network, $x \subseteq R^{3 \times tH \times tW}$ is the high-resolution image result output by the network. $d(\cdot)$ is a down-sampling operator, f_θ is the parameter of the neural network. It adjusts the size of the image by factor t, t is the up-sampling factor, and it downsamples high-resolution images to the same size as low-resolution images.

3.2 Segmentation Entropy

In the previous section, we train the DIP network by inputting the random coding vector z, and the network output the high-resolution images in each

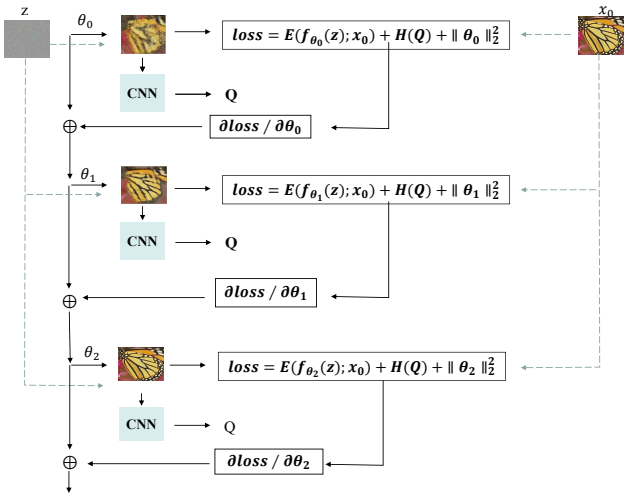

Fig. 4 This is the specific flow chart of each iteration optimization in the joint training. The random vector z is used as input, and the high resolution image is generated after being sent to the DIP network. Then we send the high resolution image to the segmentation network to obtain the segmentation entropy. We add segmentation entropy to each iteration of DIP to guide the generation of high resolution images. From this, we can see that the segmentation entropy also plays a role in the DIP network parameters.

iteration. The obtained high-resolution images are sent to the segmentation network for training. The specific process is as follows.

Firstly, using a regular grid of size $h \times w$ to partition the $H \times W$ image, and consider each grid cell as an initial superpixe. Here h, w are the hight and width of the superpixel block. H, W are the hight and width of the image. This step is to initialize the superpixel center. To get the final superpixel segmentation map, we need to find each pixel $p = (u, v)$ belongs to which cluster center $s = (i, j)$ by a mapping g, (u, v) is the coordinate position of the pixel, (i, j) is the coordinate position of the cluster center. If it find that the current pixel p belongs to a cluster center s, set the map $g_s(p) = g_{i,j}(u, v) = 1$. However, a pixel is only related to several clustering centers around it, and the connection with other clustering centers can be ignored. In order to reduce the calculation amount, only nine clustering centers around it are calculated. Therefore, the mapping is written as $g \subseteq Z^{H \times W \times 9}$. There are many methods to calculate mapping g, such as calculating the euclidean distance, but in this paper, we use deep neural network to predict g directly. In order to make our objective function differentiable, we use a soft correlation mapping $\mathrm{q} \in \mathrm{Z}^{H \times W \times \mathcal{N}_\mathbf{p}}$ to replace g. We emphasize that q is the predicted weight of each pixel to the nine cluster centers around the pixel, and the sum of the weights is 1. $\mathcal{N}_\mathbf{p}$ is the number of cluster centers. Here $q_s(p)$ represents the probability that each pixel p is assigned to the surrounding cluster center s, $s \subseteqq \mathcal{N}_\mathbf{p}$ and $\sum_{\mathbf{s} \in \mathcal{N}_\mathbf{p}} q_\mathbf{s}(\mathbf{p}) = 1$.

$$\mathbf{u_s} = \frac{\sum_{\mathbf{p}:\mathbf{s}\in\mathcal{N}_\mathbf{p}} \mathbf{f}(\mathbf{p}) \cdot q_\mathbf{s}(\mathbf{p})}{\sum_{\mathbf{p}:\mathbf{s}\in\mathcal{N}_\mathbf{p}} q_\mathbf{s}(\mathbf{p})}, \mathbf{l_s} = \frac{\sum_{\mathbf{p}:\mathbf{s}\in\mathcal{N}_\mathbf{p}} \mathbf{p} \cdot q_\mathbf{s}(\mathbf{p})}{\sum_{\mathbf{p}:\mathbf{s}\in\mathcal{N}_\mathbf{p}} q_\mathbf{s}(\mathbf{p})} \tag{6}$$

The superpixel correlation matrix q is predicted by using the standard encoder-decoder design with skip connection. Firstly, the encoder takes the high-resolution image generated by DIP as the input and generates the high-level feature map through the convolutional neural network. Then, the decoder gradually samples the feature map through the deconvolution layer to predict the correlation matrix q finally. In terms of loss function, f_p is the pixel attribute of pixels. In this method, f_p is the 3D CIELAB color vector. We further represent the position of the pixel by the image coordinate $p = [x, y]^T$ of the pixel, where x is the abscissa and y is the ordinate.

By the association graph q, we can predict the color and location properties of superpixel centers by Eq (6). The attributes of the superpixel center can be represented as $C_s = (U_s, I_s)$, where U_s is the CIELAB color attribute vector of the superpixel center and I_s is the location attribute vector of the superpixel center.

Eq (6) is a clustering process. A pixel will assign its own attributes to the surrounding cluster centers with a certain weight. If the predicted pixel belongs to the current cluster center, the current weight is set to be large, and then, update the properties of the cluster center.

The computation from pixels to cluster centers can not be used as a basis for superpixel segmentation. It also needs to use Eq (7) to reconstruct the original pixel attribute with the properties of the superpixel center. That is to say, the original pixel attribute is reconstructed again by the correlation matrix q and the superpixel center. The follow two steps complete the segmentation of superpixel.Firstly,the superpixel center is found by pixel clustering, and then the clustering center attribute is reconstructed to the pixel attribute.

$$\mathbf{f}'(\mathbf{p}) = \sum_{\mathbf{s}\in\mathcal{N}_\mathbf{p}} \mathbf{u_s} \cdot q_\mathbf{s}(\mathbf{p}), \mathbf{p}' = \sum_{\mathbf{s}\in\mathcal{N}_\mathbf{p}} \mathbf{l_s} \cdot q_\mathbf{s}(\mathbf{p}) \tag{7}$$

$f^{'}(p)$ is the reconstructed pixel color attribute and $p^{'}$ is the reconstructed pixel position attribute. In order to complete the training of superpixel segmentation network, a loss function as Eq (8) is designed. Loss function has two parts, one is the content reconstruction loss, choose the L_2-norm as the distance metric, constraint color attribute loss. The second is the spatial position loss, forcing superpixel to be compact in space. Here s is the superpixel sampling interval, m is to balance the weight of these two items, m is set by ourself. The training process is shown in Algorithm 2.

$$Loss_{seg} = \sum_{\mathbf{p}} \|\mathbf{f}(\mathbf{p}) - \mathbf{f}'(\mathbf{p})\|_2 + \frac{m}{S} \|\mathbf{p} - \mathbf{p}'\|_2 \tag{8}$$

Algorithm 2 Segmentation Entropy

Input: High-resolution image x
Output: A soft correlation mapping q

repeat
  $q = {\theta_s}^{(i)}(x)$
  Update superpixel center $C_s = (U_s, I_s)$ by Eq (6).
  Update the reconstructed $\mathbf{f}'(P), P'$ by Eq (7).
  Update the seg_{loss} by Eq (8).
  Update ${\theta_s}^{(i)}$ using the ADAM algorithm.
  $i \leftarrow i + 1$
until $i > i_{max}$

Algorithm 3 Super-resolution image algorithm

Input: Random code vector z, low-resolution image x_0.
Output: High-resolution image x.

initialize $x^{(0)} = {\theta_x}^{(0)}(z)$ by Eq (5).
repeat
  repeat
    $q = {\theta_s}^{(j)}(x^i)$
    update superpixel center $C_s = (U_s, I_s)$ by Eq (6).
    Update the reconstructed $\mathbf{f}'(P), P'$ by Eq (7).
    Update the seg_{loss} by Eq (8).
    Update ${\theta_s}^{(j)}$ using the ADAM algorithm.
    $j \leftarrow j + 1$
  until $j > j_{max}$
  $q = \theta_s(x^i)$
  Build $H(q)$ according to Eq (9).
  Update the *loss* by Eq (10).
  Update ${\theta_x}^{(i)}$ using the ADAM algorithm.
  $x^{(i)} = {\theta_x}^{(i)}(z)$
  $i \leftarrow i + 1$
until $i > i_{max}$

3.3 Combined Training

In the joint training process, we send the images generated by each iteration of DIP into the superpixel segmentation network. By Eq (8), we can train the superpixel segmentation network. We obtain the corresponding segmentation entropy by Eq (9), $\sum_{i=1}^{\mathcal{N}_\mathbf{p}} q_i = 1$. q_i is the probability of an event occurring, it represents the probability that the pixel p belongs to the surrounding superpixel block. $\mathcal{N}_\mathbf{p}$ is the number of cluster centers. Then, the segmentation entropy is added to the loss of DIP and perform the next iterative training. The process is shown in Fig. 3, Fig. 4 and Algorithm 3.

$$H(q) = -\sum_{i=1}^{\mathcal{N}_{\mathbf{P}}} q_i \log q_i \tag{9}$$

$$Loss = E\left(f_\theta(\mathbf{z});\ x_{\mathbf{0}}\right) + \boldsymbol{H}(\boldsymbol{q}) + \|\boldsymbol{\theta}\|_{\mathbf{2}}^{2} \tag{10}$$

4 Experimental Evaluation and Results

4.1 Dataset

In order to evaluate the proposed algorithm, we carried out experiments on common benchmark datasets, including Set5, Set14 and BSD100. The Set5 dataset contains five images, Set14 contains 14 images and BSD100 contains 100 complex natural images. These datasets are used for single image super-resolution reconstruction. In the experiments, the size of the high-resolution image was the size of the original image. The low-resolution image is downsampled according to the scaling factor, and the bicubic interpolation method is used to downsample the image. So we can get matched low resolution and high resolution images, namely l_{LR} and l_{HR}. Then, l_{LR} image is used for network training, and l_{HR} is used to evaluate the training results. The batch size of the image during the training is set to 1.

4.2 Implementation Details

Our model is implemented by PyTorch. In the DIP network, the number of iterations is set to 2000. Optimizer was used Adam with $\beta 1 = 0.9$, $\beta 2 = 0.999$. The learning rate is initialized to 0.0005. Firstly, downsampling the high resolution image size, so that the size meets the size of the segmentation network. In the segmentation network, optimized using Adam with $\alpha 1 = 0.9$ and $\alpha 2 = 0.999$. We use $loss_{seg}$ in Eq (8) as the segmentation loss function, where m = 0.003. For the number of superpixels, the size of superpixel cell is 4×4, this size determines the number of superpixel cluster centers. For total training, we use Eq (10) as the loss function.

4.3 Evaluation Metrics

The evaluation criteria of single-image super-resolution methods are usually divided into subjective evaluation and objective evaluation. Subjective evaluation is to visually compare the original image with the generated image by human eyes. In order to verify the quality of the model, objective evaluation criteria such as peak signal-to-noise ratio (PSNR) and structure similarity (SSIM) are usually used to evaluate the reconstruction quality of the generated image for different models. The peak signal-to-noise ratio measures the quality of image reconstruction by calculating the error between the corresponding pixels. The higher the value is, the stronger the repair ability of the network is.

Type	Method		Set5	Set14	BSD100
			PSNR/SSIM	PSNR/SSIM	PSNR/SSIM
Supervised	LapSRN [32]	2	37.52/0.9591	33.08/0.9130	31.08/0.8950
	VDSR [33]	2	37.53/0,9590	33.05/0.9130	31.90/0.8960
	EDSR [34]	2	38.11/0.9602	33.92/0.9150	32.32/0.9013
	SRCNN [15]	2	36.66/0.9542	32.45/0.9067	31.36/0.8879
	RCAN [35]	2	38.27/0.9614	34.12/0.9216	32.41/0.9027
	SAN [19]	2	38.31/0.9620	34.07/0.9213	32.42/0.9028
Unsupervised	Bicubic	2	33.66/0.9299	30.24/0.8688	29.56/0.8431
	DIP [26]	2	32.43/0.9039	29.07/0.8509	28.48/0.8170
	Ours	2	32.95/0.9097	29.56/0.8543	28.75/0.8241
Supervised	LapSRN [32]	4	31.54/0.8850	28.19/0.7720	27.32/0.7270
	VDSR [33]	4	31.35/0.8830	28.02/0.7680	27.29/0.7260
	EDSR [34]	4	32.46/0.8968	28.80/0.7876	27.71/0.7420
	SRCNN [15]	4	30.48/0.8628	27.50/0.7513	26.90/0.6675
	SAN [19]	4	32.64/0.9003	28.92/0.7888	27.78/0.7436
	RCAN [35]	4	32.63/0.9002	28.87/0.7889	27.77/0.7436
Unsupervised	Bicubic	4	28.42/0.8104	26.00/0.7027	25.96/0.6675
	DIP [26]	4	29.68/0.8495	26.87/0.7412	26.35/0.7053
	Ours	4	29.98/0.8532	27.01/0.7433	26.40/0.7062
Supervised	LapSRN [32]	8	26.15/0.7380	24.35/0.6200	24.54/0.5860
	VDSR [33]	8	25.93/0.7240	24.26/0.6140	24.49/0.5830
	EDSR [34]	8	26.96/0.7762	24.91/0.6420	24.81/0.5985
	SRCNN [15]	8	25.33/0.6900	23.76/0.5910	24.13/0.5660
	SAN [19]	8	27.22/0.7829	25.14/0.6470	24.88/0.6011
	RCAN [35]	8	27.31/0.7878	25.23/0.6511	24.98/0.6059
Unsupervised	Bicubic	8	24.40/0.6580	23.10/0.5660	23.67/0.5480
	DIP [26]	8	25.88/0.7120	24.11/0.6079	24.27/0.5707
	Ours	8	25.98/0.7131	24.12/0.6083	24.30/0.5710

Table 1 Comparison with existing methods on Image super-resolution tasks. We compared the PSNR and SSIM of these method under Set5, Set14 and BSD100 with three scale factors of 2, 4 and 8.

4.4 Experimental Results

The experiment in this paper is to compare the reconstructed images by our method with the images generated by other different methods. We compare our method with the existing supervised method and unsupervised method on the Set5, Set14 and BSD100 datasets at three different scale factors. Our method is to drive the network to produce better edge details through the segmentation entropy obtained by image segmentation. As shown in Table 1, By adding the segmentation entropy, our method achieves better results than DIP when the scale factor is 2, 4. When the scale factor is 8, the improvement effect is not obvious. This is because the image generated by the DIP network has clear image edges when the scale factor is not large. The clearer the image edge, the lower the segmentation entropy. Clearer image edges can force the DIP to produce clearer images. Similarly, with the increase of scale factor, the

Method	Set5			Set14		
	×2	×4	×8	×2	×4	×8
	PSNR/Entropy	PSNR/Entropy	PSNR/Entropy	PSNR/Entropy	PSNR/Entropy	PSNR/Entropy
DIP	32.43/115369	29.68/119723	25.88/204765	29.07/236056	26.87/248467	24.11/440273
Ours	32.95/111063	29.98/107942	25.97/119848	29.56/222684	27.01/204480	24.12/216402

Table 2 Segmentation entropy generated by our method and DIP method on Set5 and Set14 dataset. We send the image obtained by our method and the image obtained by DIP into the segmentation network, and use Eq (9) to calculate the segmentation entropy. Experimental results show that the segmentation entropy generated by our method is smaller.

cell size	×2	×4	×8
	PSNR/SSIM/Entropy	PSNR/SSIM/Entropy	PSNR/SSIM/Entropy
4×4	32.95/0.9097/111063	29.98/0.8532/107942	25.97/0.7062/119848
8×8	32.98/0.9099/100512	30.01/0.8535/102558	25.98/0.7063/103323
16×16	33.04/0.9011/86331	30.07/0.8540/88743	26.03/0.7065/88954

Table 3 We compared the effect of the number of cluster centers on the experimental results under different scale factors in the set5 dataset. The cell size determines the number of superpixel cluster centers. The larger the size is, the less the number of cluster centers is.

scale factor	Nearest	Bicubic	Area
	PSNR	PSNR	PSNR
×2	30.24	32.63	32.89
×4	24.91	28.92	29.52
×8	21.37	24.95	25.54

Table 4 The influence of different down-sampling methods on the results of DIP.

edge of the obtained image becomes more and more blurred, resulting in the increase of segmentation entropy, and the energy can not flow into DIP well.

As shown in Fig. 5 and Fig. 6, we compare the subjective effects of our method with other different methods. The function of the segmentation entropy is to make each pixel better divided into the superpixel block in each iterative training, so that the edge details of the image can be better recovered. Fig. 5, Fig. 6 show that Our method can produce clearer image edges than DIP by adding segmentation entropy.

However, supervised methods use pairs of data for supervised training, so the edges of the image will be clearer, while our method does not use pairs of data and only uses an image for image restoration. Our method only restores the image through the image prior inside the image, lacking the supplement of external data.

Table 2 is the segmentation entropy results of images trained by DIP and our method on Set5 dataset. The results show that our method can produce lower segmentation entropy on each image. This also confirms why our method can get a clearer edge.

In Table 3, we compare the influence of the number of cluster centers on the experimental results on the Set5 dataset. We adjust the number of cluster centers by changing the size of superpixel cell. The larger the size, the smaller the number of cluster centers. From Table 3, we can see that the fewer the number of cluster centers, the smaller the segmentation entropy, and the greater the PSNR value. The reason for this phenomenon is that the fewer the number of clustering centers, the greater the weight difference between each pixel and the surrounding clustering centers, and finally the smaller the segmentation entropy obtained by training. The smaller segmentation entropy can drive DIP to produce clearer image edges.

At the same time, we compare the effects of different DIP downsampling methods on the results when the scale factors are 2 , 4 and 8. In the area interpolation method, bicubic and nearest interpolation method, the image effect obtained by the area interpolation method is the best. Therefore, among the three methods, the image effect obtained by joint training using the area interpolation method is the best. This shows that DIP using better interpolation method, the final joint training can get better image.

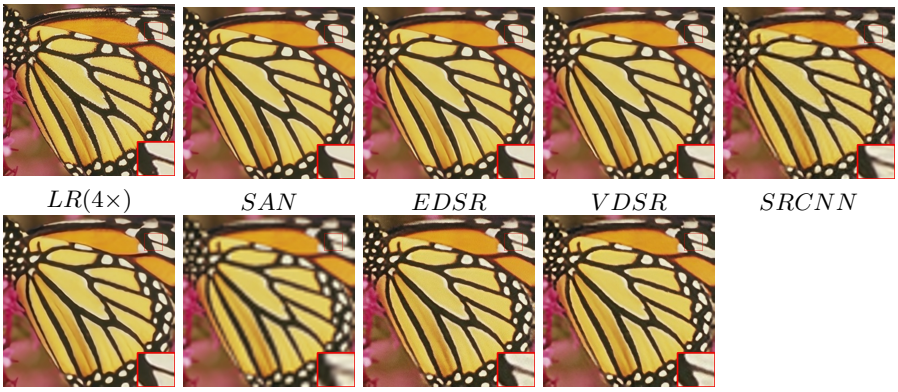

Fig. 5 Subjective results. The image is the third image in the Set5 dataset.

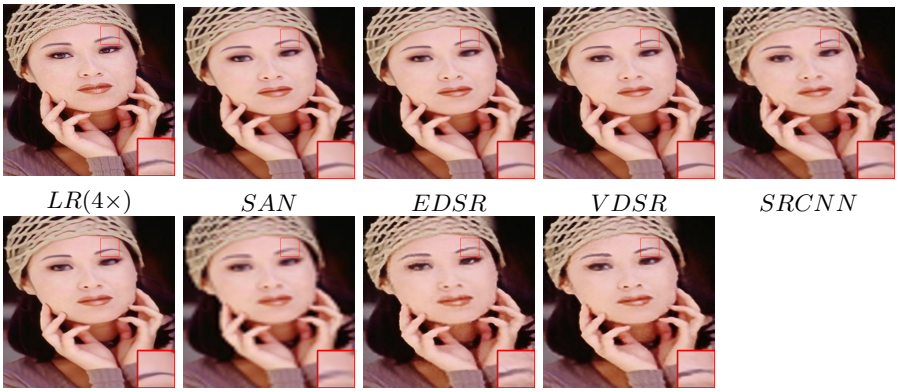

Fig. 6 Subjective results. The image is the fifth image in the Set5 dataset.

5 Conclusions

In this paper, we propose an unsupervised deep image SR method that based on segmentation driven. This method does not need pre-training and does not

depend on a large number of datasets and GT, only an image is needed for image super-resolution. Our approach is completely unsupervised. The image can be restored using only an image. DIP implicitly uses the network architecture to obtain the regularization effect of the restored image. By adding the segmentation driven to DIP, it will provide additional improvement. We use back propagation to inject the gradient generated by segmention entropy energy into DIP to obtain lower energy optimization parameters. Adding the segmention entropy will force the restored image to have clear boundary. Experiments show that the proposed method can obtain more abundant experimental results of edges and details.

Our method can produce clearer image edges than DIP, but it can not be improved well when the scale factor is too large. The disadvantage is that we need to train the segmentation network in each iteration of DIP, and the training time is long. And our method is not very ideal when the scale factor is too large.

References

[1] Greenspan, H.: Super-resolution in medical imaging. The computer journal **52**(1), 43–63 (2009)

[2] Li, Y., Sixou, B., Peyrin, F.: A review of the deep learning methods for medical images super resolution problems. Irbm **42**(2), 120–133 (2021)

[3] Mahapatra, D., Bozorgtabar, B., Garnavi, R.: Image super-resolution using progressive generative adversarial networks for medical image analysis. Computerized Medical Imaging and Graphics **71**, 30–39 (2019)

[4] Dong, R., Zhang, L., Fu, H.: Rrsgan: Reference-based super-resolution for remote sensing image. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–17 (2021)

[5] Lei, S., Shi, Z., Zou, Z.: Super-resolution for remote sensing images via local–global combined network. IEEE Geoscience and Remote Sensing Letters **14**(8), 1243–1247 (2017)

[6] Merino, M.T., Nunez, J.: Super-resolution of remotely sensed images with variable-pixel linear reconstruction. IEEE Transactions on Geoscience and Remote Sensing **45**(5), 1446–1457 (2007)

[7] Cao, S., Wu, C.-Y., Krähenbühl, P.: Lossless image compression through super-resolution. arXiv preprint arXiv:2004.02872 (2020)

[8] He, C., Liu, L., Xu, L., Liu, M., Liao, M.: Learning based compressed sensing for sar image super-resolution. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **5**(4), 1272–1281 (2012)

[9] Sen, P., Darabi, S.: Compressive image super-resolution. In: 2009 Conference Record of the Forty-Third Asilomar Conference on Signals, Systems and Computers, pp. 1235–1242 (2009). IEEE

[10] Blu, T., Thévenaz, P., Unser, M.: Linear interpolation revitalized. IEEE Transactions on Image Processing **13**(5), 710–719 (2004)

[11] Li, X., Orchard, M.T.: New edge-directed interpolation. IEEE transactions on image processing **10**(10), 1521–1527 (2001)

[12] Nayak, R., Patra, D.: Image interpolation using adaptive p-spline. In: 2015 Annual IEEE India Conference (INDICON), pp. 1–6 (2015). IEEE

[13] Schultz, R.R., Stevenson, R.L.: A bayesian approach to image expansion for improved definition. IEEE Transactions on Image Processing **3**(3), 233–242 (1994)

[14] Sun, J., Xu, Z., Shum, H.-Y.: Image super-resolution using gradient profile prior. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2008). IEEE

[15] Dong, C., Loy, C.C., Tang, X.: Accelerating the super-resolution convolutional neural network. In: European Conference on Computer Vision, pp. 391–407 (2016). Springer

[16] 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, pp. 4681–4690 (2017)

[17] 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, pp. 1874–1883 (2016)

[18] 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), pp. 286–301 (2018)

[19] Dai, T., Cai, J., Zhang, Y., Xia, S.-T., Zhang, L.: Second-order attention network for single image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11065–11074 (2019)

[20] Bulat, A., Yang, J., Tzimiropoulos, G.: To learn image super-resolution, use a gan to learn how to do image degradation first. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 185–200 (2018)

[21] Yuan, Y., Liu, S., Zhang, J., Zhang, Y., Dong, C., Lin, L.: Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 701–710 (2018)

[22] Maeda, S.: Unpaired image super-resolution using pseudo-supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 291–300 (2020)

[23] Wei, Y., Gu, S., Li, Y., Timofte, R., Jin, L., Song, H.: Unsupervised real-world image super resolution via domain-distance aware training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13385–13394 (2021)

[24] Fritsche, M., Gu, S., Timofte, R.: Frequency separation for real-world super-resolution. In: IEEE/CVF International Conference on Computer Vision (ICCV) Workshops (2019)

[25] Zhou, Y., Deng, W., Tong, T., *et al.*: Guided frequency separation network for real-world super-resolution. In: CVPR Workshops (2020)

[26] Ulyanov, D., Vedaldi, A., Lempitsky, V.: Deep image prior. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9446–9454 (2018)

[27] Wang, J., Shao, Z., Huang, X., Lu, T., Zhang, R., Ma, J.: Enhanced image prior for unsupervised remoting sensing super-resolution. Neural Networks (2021)

[28] Shocher, A., Cohen, N., Irani, M.: "zero-shot" super-resolution using deep internal learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3118–3126 (2018)

[29] Soh, J.W., Cho, S., Cho, N.I.: Meta-transfer learning for zero-shot super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3516–3525 (2020)

[30] Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I., Frey, B.: Adversarial autoencoders. arXiv preprint arXiv:1511.05644 (2015)

[31] Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)

[32] 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, pp. 624–632 (2017)

[33] 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, pp. 1646–1654 (2016)

[34] Lim, B., Son, S., Kim, H., Nah, S., Mu Lee, K.: Enhanced deep residual networks for single image super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 136–144 (2017)

[35] 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), pp. 286–301 (2018)