Dynamic Unstructured Pruning Neural Network Image Super-Resolution Reconstruction

Shaowei Zhang[1], Rongwang Yin[2], and Mengzi Zhang[1*]
[1]Department of Computer Engineering, Anhui Wenda University of Information Engineering, Hefei 231201, China.
[2]Basic Experimental and Practical Training Center, Hefei University, Hefei 230601, China.
E-mail: mzzhang118@163.com
*Corresponding author

Numerous images based on deep learning techniques for super-resolution reconstruction increase the network capacity to express features mainly by deepening the network. However, excessively extending the network's depth causes the model to be over-parameterized and complicated. Furthermore, redundant parameters increase the instability of feature expression. To address this issue, based on the unstructured pruning algorithm, the weight parameters are changed and the balanced learning strategy is used, this paper proposes a neural network unstructured pruning algorithm that is suitable for image super-resolution reconstruction tasks, called the unstructured pruning algorithm. Without changing the network structure and increasing the computational complexity, the overall feature expression ability of the network is improved by searching for an optimal yet sparse sub-network of the original network, which excludes the influence of redundant parameters and maximizes the ability to capture fine-grained and richer features with limited parameters. Based on the Set5, Set14, and BSD100 test sets, experimental outcomes demonstrated that compared with an original network model and unstructured pruning algorithm, SSIM and PSNR of the rebuilt images obtained by Dynamic unstructured pruning algorithm are improved, and they have richer detail features and clearer overall and local contours.

Povzetek: Predstavljeno je dinamično neurejeno obrezovanje (krčenje) nevronske mreže za naloge rekonstrukcije - obnovitev večje resolucije slik. Eksperimentalni rezultati na testnih zbirkah Set5, Set14 in BSD100 kažejo izboljšave v SSIM in PSNR merah.

1 Introduction

Super-resolution reconstruction technology for single-frame images is extensively utilized in numerous domains, like visual imaging, monitoring images, satellite imaging for remote sensing, and medical imaging, etc. of mobile devices. The two primary categories of available SISR reconstruction techniques are depth learning-based and interpolation-based. Interpolation by reconstruction algorithm has low computational complexity and fast reconstruction speed. It inserts a number of pixels at appropriate locations according to the known feature information of low-resolution images and spatial dimensional correlation to improve the resolution of low-resolution images. However, such algorithms mainly expand the feature information of low-resolution images by simply evaluating the correlation between adjacent pixels, Therefore, the method of reconstruction, the detailed features of high-resolution images are lost due to an issue in obtaining the high-frequency information, which also produces an undesirable visual effect in the imaging. The depth-learning-based reconstruction algorithm uses a neural network model to determine the relationship between the respective high-resolution and low-resolution images through mapping. The low-resolution image is subsequently subjected to the mapping relationship and uses previous information to map it to the high-quality high-resolution image. Research on Reference [1-3] shows that the algorithm based on depth learning has a more powerful reconstruction performance than the algorithm based on interpolation, and can greatly enhance the image the reconstruction quality.

In reconstruction algorithm according to depth learning, the network layer with strong feature expression ability can learn the image feature representation more effectively, and capture more fine-grained and rich details. Some SISR deep learning-based reconstruction methods [3-4] enhance the network model's overall capability to convey features by using deeper network layers and more complex connection methods, thereby improving the reconstruction performance. Although the overall feature expression ability of the network layer that enhanced by expanding the network's depth layer and complex connection methods, excessive dependence on this method will lead to a sharp rise in the parameters number and The network model's computational difficulty. In the neural network,

the parameters represent the network scale model, and the computational complexity represents the efficiency of forward reasoning of the network model. In the sense of practical application, the applicability of the SISR network model in real circumstances will be constrained by the network model's scale and the speed of super-resolution image reconstruction. Reference [5] pointed out that discarding redundant parameters in the deep network layer can not only reduce the negative impact of over-parameterization of the model of network but enhance the feature expression's stability.

The neural network's pruning mechanism lowers the network model's parameters by eliminating some of variables. Structured and unstructured pruning techniques are the two categories into which existing pruning techniques can be separated. A structured pruning algorithm reduces the number of parameters by discarding some channels in the network layer. Because this method is an adjustment of the network structure, the original network structure will change. The unstructured pruning algorithm discards some parameters by zeroing some parameters in the network layer, so the original network structure will not be changed. Reference [6-7] applies the structured pruning algorithm to the image super-resolution model of network and maximizes the network size representation to deploy it to the actual application scenarios while ensuring the reconstruction performance.

Reference [8] proposed an unstructured pruning algorithm (UPru) for neural networks. The design initiative of the algorithm to treat the complex neural net as the prize pool, and the winning lottery is a sparse sub-network corresponding to a group of weight parameters. The outcomes of the experiment demonstrate the validity of UPru algorithm has achieved remarkable resulted in image classification tasks by searching the optimal sparse sub-network, but it is not effective in image super-resolution tasks.

UPru algorithm adopts an unbalanced feature learning strategy, which only focuses on the sparsity of the network model and ignores the diversity of feature expression. To enhance the effectiveness of image reconstruction, this paper indicates the dynamic unstructured pruning algorithm (DUPru) according to balanced learning strategy in combination with the characteristics of an image super-resolution reconstruction task. This algorithm not only ensures the sparsity of the network model, but also pays attention to the diversity of weight parameter learning, which can solve the problem of poor reconstruction performance caused by over-parameterization of the network model in the high-resolution image reconstruction assignment, and enhance the super-resolution image's reconstruction quality without changing the network structure and increasing the computational complexity.

2   Related Work

2.1   Depth learning-based technique for reconstructing super-resolution images from single frames

Reference [9] proposed SRCNN model, which realized the low-resolution feature extraction images, nonlinear mapping features, and reconstruction of high-resolution images through three layers of CNN. Compared with the traditional interpolation-based algorithm for super-resolution restoration of images, the SRCNN reconstructed high-quality image has more abundant details, and the contour of the image is visible. Reference [10] proposed a DRN network model to reduce the dependence of low-resolution images on high-quality images by training the two-way mapping link between high- and low-resolution images, and then solving the super-resolution problem of real samples. Reference [11] proposed the RFANet network model, using the residual module to improve the spatial attention module's efficiency in extracting features and integrating them into the residual feature aggregation framework, to enhance the quality of images with greater clarity. The TTSR network model proposed in [12] is a learnable texture extractor, which obtains the texture information most suitable for super-resolution reconstruction through training, provides rich texture basic information for texture migration and texture synthesis, and finally generates high-quality super-resolution images.

Most of the existing SISR reconstruction algorithms enhance the feature extraction ability by designing a deeper network structure and using complex network connection strategies to learn and capture rich texture features from low-resolution images using effective feature extraction capabilities, and then obtain high-quality super-resolution images. However, this excessive expansion of network depth and the use of complex connection methods will cause the network model to grow significantly in both size and computing complexity.

2.2 UPru algorithm of neural network

UPru algorithm of neural network obtains sparse subnetworks by resetting some parameters to zero [13-15]. Generally, the sparsity of neural networks can improve the ability of feature selection and generalization. On the one hand, some researchers obtain sparse subnetworks by exploring effective UPru methods, such as using different regularization techniques or designing feasible pruning strategies. The method proposed in [13-14, 16] is to obtain sparse subnetworks by using L2 regularization technology to optimize the model based on a convolutional neural network. Reference [15] obtained sparse subnetworks through L0 regularization technology. Reference [17] realized the pruning process by combining regularization technology and parameter sensitivity evaluation. Different from the above method using regularization technology, Reference [8] searches for the optimal sparse subnet by

resetting unimportant parameters to zero through an iterative dynamic pruning process. On the other hand, some researchers pay attention to the efficiency of UPru methods and explore how to implement this efficiency on hardware devices. Reference [18] explored how to find balanced sparsity in CNN models and accelerate the reasoning process of neural networks on hardware devices. Reference [19] realized the pruning and restoration process of parameters through the evaluation of network parameters and tried to maximize the compression of models based on dense neural networks or CNN to speed up the training process. In addition, Reference [13] deployed sparse networks to separately designed hardware devices and achieved very high acceleration efficiency. Reference [25] demonstrates aligned structured sparsity learning (ASSL), which modifies the sparsity scale parameters using L2 regularization and involves a weight normalization layer. Regarding improvements to performance in terms of numbers and graphics, ASSL outperforms recent techniques. Reference [26] developed the Global Aligned Structured Sparsity Learning (GASSL). ASSL and Hessian-Aided Regularization (HAIR) are the two main parts of GASSL. Comprehensive outcomes indicate GASSL's benefits across other contemporary alternatives. Reference [27] offer progressively structured sparsity in hardware-friendly Scalable super-resolution (SR) (HSSR). The model performs more effectively, flops less, and is smaller than the Slimmable technique. Experimental results illustrate that in real-world applications, HSSR produces a significant reduction in comparison to other models. Reference [28] develops SLS, an innovative layer-wise pruned ratio search framework designed with N: M sparsity in mind. Our results produce a modern restoration of image performance at comparable computational budgets when compared to the earlier approaches with uniform N: M sparseness at all layers. Reference [29] described the Scalable SR framework that is memory-friendly, known as MSSR. The pruning-out method and the SR model mask that create nested set are gradually reduced by the use of LTH applying weights for rewinding. The efficacy of MMSR is demonstrated by numerous tests. Outperforming the evaluated lightweight SR methods, 94% sparsity can be attained by the smallest scale sub-network as shown in Table 1.

Table 1: Related works

Reference	Proposed	Result	Limitations
[8]	Their offered algorithm's primary difficulty is Sparse representation for single-image high resolution.	To illustrate why the recommended method is superior, they have contrasted those results with those of the contemporary bicubic interpolation, and super-resolution techniques according to sparse representation.	As arbitrary symbols are used in the basis and coefficient matrices, subtractions can occur when there are negative coefficients and bases.
[13]	To enable to combination of the best aspects of Anchored Neighborhood Regression (ANR) with SF, they provide A+, an improved form of ANR.	They achieve exceptional temporal complexity and enhanced quality, making An optimal dictionary-based super-resolution technique available at that period.	The ensuing memory needs and time complexity can represent significant barriers to these technologies' effectiveness.
[14]	They demonstrated how well the Image Net-trained model generalizes to other datasets.	They employed an alternate loss function that allowed for more than one object per image, so their performance could be improved.	The study can have limited the knowledge of how models make decisions because it does not thoroughly explore interpretability methods for convolutional networks outside of visualization.
[15]	Deep convolutional generative adversarial networks (DCGANs) are an innovative class of CNNs that the research	It gave strong support that the deep convolutional adversary pair learns, in the discriminator and	To solve the issue of instability, further effort is required.

	described.	generator, a sequence of representation from object components to scenes.	
[16]	The recommendation is to employ deep learning to achieve super-resolution (SR) on a single image.	To attain performance and speed tradeoffs, they investigate various network topologies and parameter configurations.	To deal with various raising factors, one may appear at a network.
[17]	They demonstrate that the conventional code sparse technique still adequately captures domain expertise.	Compared to existing techniques, the model significantly outperformed another method by an extensive number of images.	when sparse code can be useful, they will use the SCN method.
[18]	They recommend using a deeply recursive convolutional network (DRCN) to implement the super-resolution (SR) technique for images.	Gradients that explode or disappear make it extremely difficult to train a DRCN using a typical gradient descent method.	To employ image-level information, one can attempt more recursions.
[19]	Dynamic network surgery is the innovative network compression technique they offer.	It performs more effectively than the existing pruning method, consequently compressing the number of parameters of the LeNet-5 algorithm and AlexNet by an aspect without sacrificing accuracy.	Fewer epochs are required because the technique has a superior learning efficiency.
[25]	The article provided acquiring with aligned structured sparsity (ASSL).	ASSLN outperforms more modern techniques as a result of both quantitative and visual performance gains.	Network prune is another prevalent model compression, due to the large number of residual connections in SR, it can be difficult to train lightweight SR networks directly.
[26]	They offer Global Sparsity Learning with Structured Alignment (GASSL).	Comprehensive results demonstrate that GASSL is comparable to other variations.	The complex model architecture and rigorous training process of that model can make it difficult to scale to incredibly huge datasets and computer resources.
[27]	They offer progressively structured sparsity in hardware-friendly Scalable SR (HSSR).	Experimental results illustrate that in real-world applications, HSSR produces a significant reduction in comparison to another model.	Its applicability can be limited by its incapacity to generalize to different hardware architectures.
[28]	We develop SLS, an innovative layer-wise pruned ratio search framework designed with N: M sparsity in mind.	The results produce a modern restoration of image performance at comparable computational budgets	Using enhanced restoration performance, the adaptive inference technique makes it easier to adjust the computational limits in

		when compared to the earlier approaches with uniform N: M sparseness at all layers.	detail.
[29]	The research describes a Scalable SR framework that is memory-friendly, known as MSSR	The research evaluated lightweight SR methods and Sparsity of 94% can be reached by the smallest scale sub-network.	Using light-weight SR methods is growing in popularity as a superior solution for devices with limited resources.

3 UPru algorithm for balanced learning

3.1 UPru algorithm

The UPru algorithm of neural networks deals with the network layer's weight parameters. The objective of this process is to eliminate weight parameters that are superfluous in the networking layer or that little affect the network model's ultimate output. Therefore, this pruning algorithm only deals with the network layer's weight parameters and does not affect the specific structure of the entire network model. UPru algorithm realizes the pruning process by returning meaningless weight parameters to zero in each iteration through iterative training, to further search the optimal sparse subnet. Specifically, the UPru algorithm judges whether a weight parameter in the network layer is meaningful by comparing the size relationship between the weight parameter and the threshold. When the weight parameter's value is smaller than the threshold amount and it is considered that the weight parameter is meaningless or redundant, otherwise, it is considered that the weight parameter has learning potential and significance. threshold λ is a dynamic value, and its calculation formula is:

$$\lambda = F[f_{rank}(W^t)\ /p] \qquad (1)$$

Where W^t represents the weight parameter in any network layer after t iterations. f_{rank} is a function to sort W^t incrementally. F is a function to calculate the p-th percentile of the ordered weight parameter. For the pruning process of returning the weight parameter to zero, the UPru algorithm is implemented by multiplying the mask m by the element at the corresponding location of W^t. The definition of mask m can be expressed as:

$$\begin{cases} m^t_{(j,i,k)} = 0, |\ W^t_{(j,i,k)}\ | < \lambda \\ m^t_{(j,i,k)} = 1, |\ W^t_{(j,i,k)}\ | \ge \lambda \end{cases} \qquad (2)$$

Where i, j, and k represent the element index in a tensor. It can be seen from equation (2) that when $|\ W^t_{(i,j,k)}\ |$ is lower than the dynamic threshold λ Set the value of the element to 0 at the index position corresponding to the mask m, otherwise set the value of the element to 1. In this way, meaningless weight parameters are discarded in each round of iterative pruning, while potential weight parameters are retained to become knowledge about data feature representation. The zeroing weight parameters in the network layer can be expressed as:

$$W^t = W_0 \odot m^t \qquad (3)$$

Where W_0 is the weight parameter when the network layer performs random initialization; The $\odot$ operator represents the multiplication of elements at the same position between two tensors. By formula (3), the weight parameter W^t at the time of network layer initialization is used to initialize the weight parameter of this iteration in each round of iterative pruning, and it is taken as the initial state of model fine-tuning. Although the UPru algorithm obtains outstanding performance in image classification tasks by searching the optimal sparse sub-network, it cannot achieve good results in image super-resolution reconstruction tasks.

3.2 DUPru algorithm

The unstructured pruning method DUPru is proposed in this article and is based on the balanced learning strategy by changing the weight parameters of the unbalanced learning strategy based on the UPru algorithm. By monitoring the change process of network layer weight parameters in the training process, the weight parameters whose parameter values fall into a local small range are added to the frozen queue. Once the weight parameter is added to the freezing queue, it will keep the current value in this iteration and will not be updated. In other words, when the weight parameter becomes a minimum value in the training process, its impact on the output characteristic graph is negligible. Therefore, for such weight parameters, we can choose not to update them, but focus on the potential weight parameters. In the specific implementation, this paper achieves this by controlling the gradient generated by the weight parameter in the training process, where the gradient control can be expressed as:

$$\begin{cases} g^t_{(i,j,k)} = \dfrac{\partial L_{\text{loss}}\left(w^t_{(j,i,k)}\right)}{\partial w^t_{(i,j,k)}}, \left|W^t_{(i,j,k)}\right| \geq \text{EPS} \\ g^t_{(j,i,k)} = 0, \ \left|W^t_{(j,i,k)}\right| < \text{EPS} \end{cases} \quad (4)$$

Where: L_{loss} is the loss function; $g^t_{(i.i.k)}$ is the gradient value of weight parameter in the relevant index position in the network layer. EPS is a fixed positive threshold. In the process of training, the gradient of weight parameters whose absolute values are within the threshold range is reset to zero, which limits their learning feature expression throughout the process of improving training. Finally, the parameter optimization process of W^t can be expressed as:

$$W^t_{(j,i,k)} = W^t_{(j,i,k)} - \alpha g^t_{(j,i,k)} \quad (5)$$

Where α is the learning rate of parameter optimization. Through this balanced learning method, the DUPru algorithm proposed can maximize the learning of information about the features of the image by the weight parameters of the network layer under the premise of satisfying feature diversity. In contrast, although the UPru algorithm can maximize the sparsity of the network layer through unbalanced learning strategies, it ignores the proportion of negative order weight parameters in feature diversity learning. For image super-resolution reconstruction tasks, feature diversity learning plays the most important role.

The DUPru algorithm's procedure is explained in the following algorithm 1, where T represents the number of times of iterative pruning. E signifies the quantity of iterations. Figure 1 demonstrated the Diagrammatic representation of the dynamic unstructured pruning algorithm
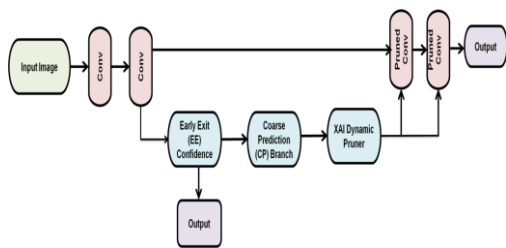

Figure 1: Diagrammatic representation of the dynamic unstructured pruning algorithm

Algorithm 1:   DUPru

Input: randomly initialized neural network model $M = f(W_0)$, mask $m = \{0,1\}^{|W|}$

(1) for $t = 1$ in T do

(2) if $t > 1$ then Calculate the p-th percentile λ of $\{|W^{t-1}| \neq 0\}$

(3) if $|W^{t-1}_{(i,j,k)}| < \lambda$ then $m^t_{(i,j,k)} = 0$

(4) Reinitialize network parameters: $W^t = W_0 \odot m^t$

(5) for $e = 1$ in E and d in D do

(6) Forward propagation: $f(d, W^t)$

(7) Calculation gradient: $g^t = \dfrac{\partial L}{\partial w^t}$

(8) if $\left|W^t_{(i,j,k)}\right| < \text{EPS}$ then $g^t_{(i,j,k)} = 0$

(9) Update weight parameter: $W^t \leftarrow (W^t, g^t)$

3 .3 MSRResNet network model based on DUPru pruning algorithm

Figure 2 depicts the general architecture of the system model MSRResNet [6] that is employed in this investigation.
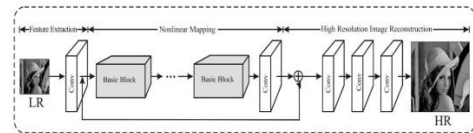

Figure 2: Framework of MSRResNet network model

The low-resolution picture ILR is initially fed into the convolution layer's feature extraction procedure as a component of the end-to-end networking model. This method can be stated as:

$$I_f = C_{ex}(I_{LR}) \quad (6)$$

Where: C_{ex} is a convolution neural network (CNN) for feature extraction. I_t is the feature map of a convolutional neural network obtains from an I_{LR} of a low-resolution image. Then, the standard diagram goes through nonlinear mapping process of the deep network module and can be expressed as:

$$I_d = C_d(I_f) \quad (7)$$

Where C_d is a convolutional neural network that realizes the nonlinear mapping of features. It is the depth feature map obtained by the feature map through the nonlinear mapping of the depth network module. Finally, the I_d is reconstructed into a super-resolution image after the process of up-sampling and feature fusion, which can be expressed as:

$$I_{HR}=C_{mer}(C_{up}(I_d)) \quad (8)$$

Where I_{HR} is the final reconstructed super-resolution image. C_{mer} is a feature fusion CNN. The cup is a CNN that realizes up sampling. One can distinguish between two types of network layers in Figure 2, first is the module of independent convolution layers as feature learning and the other is the Basic Block network module composed of multi-layer convolution layers through complex connection strategies. Among them, the Basic Block network module is an expandable and replaceable feature learning network module. It is worth noting that this paper only applies the DUPRU algorithm to the Basic Block network module in which the parameters are dominant in the network model, and uses the iterative pruning method and balanced learning strategy of this algorithm to search for the optimal sparse subnet. The particular procedure is depicted in Figure 3. In this way, redundant parameters in the Basic Block network module can be discarded, and feature learning can focus on potential weight parameters to avoid the negative impact of redundant parameters.
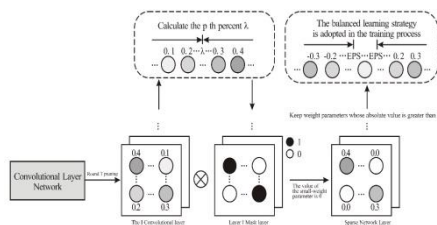

Fig. 3: The process of DUPru algorithm searching for sparse sub-networks

4 Analysis and outcomes of the experiment

4.1 Data collections and experimental parameters

To keep consistent with the previous SISR research work, this paper uses 800 training images of DIV2K [20] for training. Before training, the training data set shall be subject to data enhancement preprocessing of rotation and rotation. It's important to remember that training provides a image of the input model is randomly cropped from a high-resolution image 96×96×3 size sub-image. To ensure the reliability of the test, this paper selects Set5 [21], Set14 [22], and BSD100 [23] test data sets for experiments. At the same time, MSRA [24] is used to initialize the network parameters, Adam [25] optimizer is used to optimize them, and the initial learning rate is set to 10^{-4}. The initial parameters of the Adam optimizer are $\beta_1=0.9$, $\beta_2=0.999$, $\epsilon = 10^{-9}$。   In addition, the mean square error (MSE) loss function is used to optimize the network model.

4.2 Result analysis

This paper compares the performance of MSRResNet [6] network models using the UPru algorithm and DUPru algorithm on different test sets. To ensure the comparison's impartiality, the comparison model uses the official website code provided by the author and uses the default parameters set by the author in the experiment. The objective estimation criteria of image value, this paper use the Structural Similarity Index (SSIM) and Peak Signal Noise Ratio (PSNR) to evaluate the quality of the reconstructed super-resolution image. We analyzed the reconstruction outcomes of network model under various sparse percentages after pruning by the UPru algorithm and DUPru algorithm and further analyzed the reconstruction performance of the DUPru algorithm under different pruning rates.

4.2.1 Analysis of objective evaluation criteria

Compare the model's average SSIM and PSNR on the Set5, Set14, and BSD100 test sets, as illustrated in Tables 2 and 3. Ideal performance is indicated by a bold font. The average PSNR and SSIM of four times high-resolution images reconstructed on RGB channels using various network models are listed in Table 2.

Compared to the UPru technique, the DUPru method improves average PSNR and SSIM by 0.65 dB and 0.009 7 on the Set5 test set, 0.48 dB and 0.011 5 on a Set14 test set, and 0.37 dB and 0.011 6 on the BSD100 test set. As can be seen, the standard PSNR and SSIM for the DUPru algorithm proposed in this article are the highest on different test sets. The DUPRU technique can raise the values on Set5, Set14, and BSD100 set of tests by 0.1 dB, 0.002 1, 0.07 dB, 0.000 7, 0.08 dB, and 0.001, respectively, in comparison to the original model. The mean PSNR and the SSIM of the regenerated 4× high-resolution images on the Y channels are displayed in Table 3. It is visible that the DUPru algorithm-based network model outperforms the others on various test sets.

Table 2: Performance evaluation on the Set 14, Set 5 and BSD100 sets of tests (RGB channel)

Network	Set5		Set14		BSD100	
	PSNR /dB	SSIM	PSNR /dB	SSIM	PSNR /dB	SSIM
MSRResNet	30.13	0.863 2	26.78	0.744 7	26.21	0.711 2
MSRResNet+UPru	29.58	0.854 4	26.37	0.733 9	25.92	0.700 6
MSRResNet+DUPru	**30.23**	**0.865 3**	**26.85**	**0.745 4**	**26.29**	**0.712 2**

Table 3: Performance evaluation on Set14, Set5 and BSD100 sets of tests (Y channel)

Network	Set5		Set14		BSD100	
	PSNR /dB	SSIM	PSNR /dB	SSIM	PSNR /dB	SSIM
MSRResNet	32.02	0.892 6	28.57	0.780 8	27.54	0.734 7
MSRResNet+UPru	31.43	0.884 6	28.14	0.770 3	27.25	0.724 5
MSRResNet+DUPru	**32.12**	**0.894 3**	**28.64**	**0.781 4**	**27.59**	**0.735 4**

This article compare the reconstruction performance of network representation under different sparse percentages after the pruning process of the UPru algorithm and DUPru algorithm on the Set5 test set, as demonstrated in Figure 4. It is visible that the DUPru algorithm being used by the network model generally rises first and then decreases gradually, and obtains the optimal average PSNR when the sparse percentage is 7.95%. Although the network model using the UPru algorithm is generally similar to that using the DUPru algorithm, the algorithm uses unbalanced learning strategies and ignores the diversity of weight parameter learning, resulting in poor performance in image super-resolution tasks. In contrast, the DUPru algorithm can ensure the sparsity of network model and the diversity of weight

parameter learning. Therefore, after applying DUPru algorithm to the network model, the performance of reconstruction has been significantly improved.
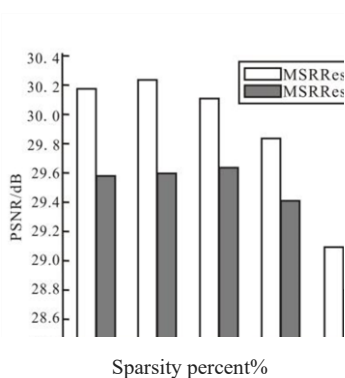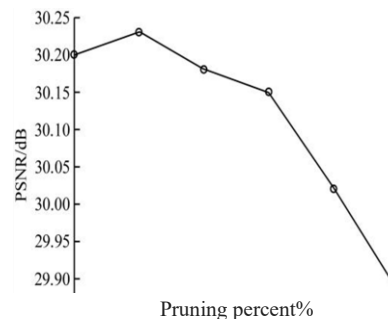

Figure 5: Comparison of PSNR under different pruning percent

Further examine the effects of the rate of pruning on network model using DUPru algorithm on the Set5 test set, as illustrated in Figure 5. The network representation can obtain a higher average PSNR under the condition of using a smaller pruning rate while using a larger pruning rate has a poor effect. The experimental results show that the network model searching for the optimal sparse subnet is a process of gradual search and fine-tuning, and cannot use a wide range of search methods.

Figure 4: Comparison of PSNR under different

4.2.2 Analysis of subjective evaluation criteria

The visual effects of the reconstructed 4× high-resolution images are compared, as demonstrated in Figure 6, and the relevant PSNR and SSIM indicators in Table 4. Figure 6 that the spots on the wings of the Butterfly image reconstructed from the model using the DUPru algorithm

are clearer and have more detailed features. The Baby image is clearer in the overall and local contours, and closer to the original image.

Table 4: Performance index for reconstructed Butterfly and Baby images

Network	Butterfly		Baby	
	PSNR/dB	SSIM	PSNR/dB	SSIM
MSRRResNet	26.93	0.902 5	28.91	0.901 5
MSRRResNet+UPrun	26.88	0.901 5	28.85	0.900 5
MSRRResNet+DUPrun	27.10	0.906 5	29.17	0.903 5
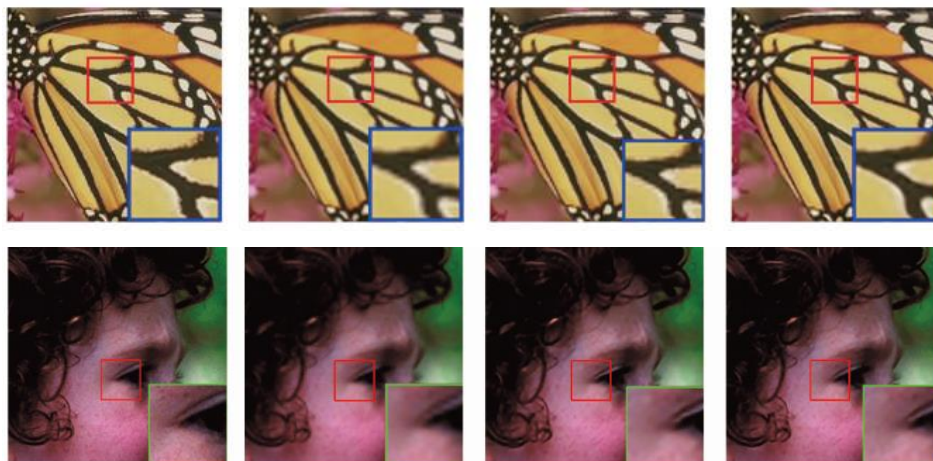
(a)MSRRResNet+DUPru   (b)MSRRResNet+UPru   (c)MSRRResNet   (d) Ground Truth

Figure 6: Comparison of the visual quality of the reconstructed Butterfly and Baby image

4.2.3 Analysis of operation efficiency

Compare the time spent in reconstructing the 4× super-resolution image in the Set5 set of test, is depicted in Table 5. To ensure fairness of comparison, this paper tests the network model in the same platform environment (Inter Core i7 11800+NVIDIA GTX2060 Super). It is demonstrated that the DUPru algorithm being used by the network model consumes the same time as the original model in super-resolution image reconstruction. The size of the network representation is consistent. Because of pruning process of the DUPru algorithm only evaluates the weight parameters through the pruning evaluation strategy in the training phase and

sets the weight parameters evaluated as redundant to zero, this unstructured pruning method only changes the size of the weight parameters in the network layer and does not change the overall structure of the network model. In addition, the DUPru algorithm plays a guiding role in the training phase, rather than a specific network layer module. Therefore, the size of the model using the DUPru algorithm is consistent with that of the innovative representation and does not increase the amount of parameters of network model. The model size is 5.2 MB. Table 6 illustrates the Comparison of various techniques effectiveness.

Table 5 Time-consuming for high-resolution image reconstructions

Image	MSRResNet	MSRResNet+DUPru
Baby	3.989 7	3.988 5
Bird	2.992 2	2.991 7
Butterfly	3.961 1	3.991 8
Head	4.988 0	3.989 5
Lena	3.989 2	3.959 7

Table 6: Comparing the effectiveness of various techniques

Methods	Set14		Set5		B100	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
SRP [30]	65	73	77	85	76	80
ASSL [31]	73	74	82	86	82	67
L_1 norm [32]	50	75	61	87	58	72
DUPru [proposesd]	97	92	94	95	98	93

4.3 Discussion

Increasing computational complexity presents issues when using ASSL [25] in picture super-resolution reconstruction. Deploying or utilizing real-time apps on resource-constrained devices is not feasible due to the complex algorithms and significant processing resources required to align organized sparsity patterns. In deep architectures in specific, DRCN [18] have issues like vanishing or expanding gradients during training. Its recursive structure can make it difficult to represent long-range relationships, which can influence the model's accuracy in reconstructing high-frequency features. Adaptability and generalization issues can occur with HSSR [27] methods. Network design can get less capable of managing a range of image properties if it is modified to accommodate hardware limitations. Our proposed algorithm performs significantly better than the existing algorithms by offering various benefits of using the proposed algorithm. It reduces overall computing complexity by optimizing model efficiency by selectively eliminating unnecessary and redundant links. Because of the resulting shorter inference times and reduced memory needs, it can be used in real-time applications and dynamic pruning allows the network to adjust by shifting input data distributions, maintaining its efficiency even with a variety of image types.

5 Conclusion

This paper proposes a dynamic unstructured pruning algorithm based on an unstructured pruning algorithm, which is suitable for high-resolution image reconstruction tasks. While ensuring sparsity of network model, it ensures the diversity of weight parameter learning features through balanced learning strategies. The outcome depicts the DUPru algorithm significantly improve the standard of the rebuilt high-resolution images on SISR network model without changing the network structure and increasing the computational complexity. Later, the structural pruning algorithm of the neural network will be applied in the image high-resolution reconstruction efforts to improve image quality and efficiency.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] H.S. Hou, H.C. Andrews, Cubic spline for image interpolation and digital filtering, IEEE Trans. Signal Process. 26 (1978) 508–517.

[2] S. Dai, M. Han, W. Xu, Y. Wu, Y. Gong, Soft edge smoothness prior for alpha channel super-resolution, in: IEEE Conference on Computer Vision and Pattern Classification (CVPR), 2007, pp. 1–8.

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

[4] R.C. Hardie, K.J. Barnard, E.A. Armstrong, Joint map registration and high resolution image estimation using a sequence of undersampled images, IEEE Trans. Image Process. 6 (1997) 1621–1633.

[5] S. Farsiu, M.D. Robinson, M. Elad, P. Milanfar, Fast and robust multiframe super-resolution, IEEE Trans. Image Process. 13 (2004) 1327–1344.

[6] M.E. Tipping, C.M. Bishop, Bayesian image super-resolution, in: Advances in Neural Information and Processing Systems 16 (NIPS), 2003.

[7] S. Baker, T. Kanade, Limits on super-resolution and how to break them, IEEE Trans. Pattern Anal. Mach. Intell. 24 (9) (2002) 1167–1183.

[8] J. Yang, J. Wright, T.S. Huang, Y. Ma, Image super-resolution via sparse representation, IEEE Trans. Image Process. 19 (11) (2010) 2861–2873.

[9] F. Yeganli, M. Nazzal, M. Unal, H. Ozkaramanli, Image super resolution via sparse representation over coupled dictionary learning based on patch sharpness, in: Proc. EMS, Prague, Czech Republic, Oct. 2014, pp. 203–208.

[10] Y. Zhang, W. Wu, Y. Dai, X. Yang, B. Yan, W. Lu, Remote sensing images super resolution based on sparse dictionaries and residual dictionaries, in: Proc. DASC, Chengdu, China, Dec. 2013, pp. 318–323.

[11] C.-H. Fu, H. Chen, H. Zhang, Y.-L. Chan, Single image super resolution based on sparse representation and adaptive dictionary selection, in: Proc. DIP, Hong Kong, Aug. 2014, pp. 449–453.

[12] W. Dong, L. Zhang, G. Shi, X. Wu, Image deblurring and super resolution by adaptive sparse domain selection and adaptive regularization, IEEE Trans. Image Process. 20 (7) (2011) 1838–1857.

[13] R. Timofte, V. De Smet, L. Van Gool, A+: Adjusted anchored neighborhood regression for fast super-resolution, in: Asian Conference on Computer Vision (ACCV), Springer, 2014, pp. 111–126.

[14] M.D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: European Conference on Computer Vision (ECCV), Springer, 2014, pp. 818–833.

[15] Radford, L. Metz, S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks, in: International Conference on Learning Representations (ICLR), 2016.

[16] C. Dong, C.C. Loy, K. He, X. Tang, Image super-resolution using deep convolutional networks, IEEE Trans. Pattern Anal. Mach. Intell. 38 (2) (2015) 295–307.

[17] Z. Wang, D. Liu, J. Yang, W. Han, T. Huang, Deep networks for image super resolution with sparse prior, in: IEEE International Conference on Computer Vision, 2015.

[18] Deeply-recursive convolutional network for image super resolution, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[19] Guo Y W, Yao A B, Chen Y R. Dynamic network surgery for efficient DNNs[EB/OL].[2021-08-09]. https://arxiv.org/pdf/1608.04493.pdf.

[20] Lugmayr A, Danelljan M, Timofte R. Ntire 2020 challenge on real-world image super-resolution: methods and results[C]//Proceedings of 2020 IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2020: 2058-2076.

[21] Bevilacqua M, Roumy A, Guillemot C, et al. Low-complexity single-image super-resolution based on nonnegative neighbor embedding[C]//Proceedings of British Machine Vision Conference. Berlin, Germany: Springer, 2019:

135.

[22] Zeyde R, Elad M, Protter M. On single image scale-up using sparse-representations[C]//Proceedings of International Conference on Curves and Surfaces. Berlin, Germany: Springer, 2020: 711-730.

[23] Martin D, Fowlkes C, Tal D, et al. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics[C]//Proceedings of 2020 IEEE International Conference on Computer Vision. Washington D. C., USA: IEEE Press, 2020: 416-423.

[24] He K M, Zhang X Y, Ren S Q, et al. Delving deep into rectifiers: surpassing human-level performance on ImageNet classification[C]//Proceedings of 2018 IEEE International Conference on Computer Vision. Washington D. C., USA: IEEE Press, 2018: 1026-1034.

[25] Zhang Y, Wang H, Qin C, Fu Y. Aligned structured sparsity learning for efficient image super-resolution. Advances in Neural Information Processing Systems. 2021 Dec 6;34:2695-706.

[26] Wang H, Zhang Y, Qin C, Van Gool L, Fu Y. Global Aligned Structured Sparsity Learning for Efficient Image Super-Resolution. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2023 Apr 20.

[27] Ye F, Lin J, Huang H, Fan J, Shi Z, Xie Y, Qu Y. Hardware-friendly Scalable Image Super Resolution with Progressive Structured Sparsity. InProceedings of the 31st ACM International Conference on Multimedia 2023 Oct 26 (pp. 9061-9069).

[28] Oh J, Kim H, Nah S, Hong C, Choi J, Lee KM. Attentive fine-grained structured sparsity for image restoration. InProceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2022 (pp. 17673-17682).

[29] Lin J, Luo X, Hong M, Qu Y, Xie Y, Wu Z. Memory-Friendly Scalable Super-Resolution via Rewinding Lottery Ticket Hypothesis. InProceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2023 (pp. 14398-14407).

[30] Zhang Y, Wang H, Qin C, Fu Y. Learning efficient image super-resolution networks via structure-regularized pruning. InInternational conference on learning representations 2021 Oct 6.

[31] Zhang Y, Wang H, Qin C, Fu Y. Aligned structured sparsity learning for efficient image super-resolution. Advances in Neural Information Processing Systems. 2021 Dec 6;34:2695-706.

[32] Li H, Kadav A, Durdanovic I, Samet H, Graf HP. Pruning filters for efficient convnets. arXiv preprint arXiv:1608.08710. 2016 Aug 31.