Improved Double Regression Nonlinear Image Super Resolution Model

Jieyi Lv
Xi'an Technological University
State and Provincial Joint Engineering Lab. of
Advanced Network, Monitoring and Control
No.2 Xuefu Middle Road, Weiyang district
Xi'an, Shaanxi, China
E-mail: ljylly150@163.com

Zhongsheng Wang
Xi'an Technological University
State and Provincial Joint Engineering Lab. of
Advanced Network, Monitoring and Control
No.2 Xuefu Middle Road, Weiyang district
Xi'an, Shaanxi, China
E-mail: wzhsh1681@163.com

***Abstract*—The existing super resolution reconstruction methods are mainly divided into traditional super resolution reconstruction and deep learning super resolution reconstruction. The main problem faced by traditional super resolution reconstruction algorithms, such as image enlargement and space transformation, is how to establish the mapping relationship between the input image and the target image, and express the pixel value of the target image through the mapping relationship. As a prominent problem, the difficulty of super resolution reconstruction lies in the fact that there is no realizable matrix relationship between one - to - many mapping relationships. Based on the U-Net network framework, this paper improves the jump-connected modules. By using the combination of convolutional layer, activation layer and residual channel block, the overall module operation efficiency is increased by 2.4%, the overall PNSR is increased by 0.49db, and the running speed is increased by 0.3ms on average when processing a single image compared with other classical models.**

Keyword—Super-resolution; Double Regression; U-Net network; Model Refinement

I. INTRODUCTION

With the continuous development of computer hardware and software technology as well as image and video sensor technology, a huge amount of image information is generated every day. How to obtain the hidden available information in the mass image is always a research topic with great value in computer vision. In recent years, the super resolution reconstruction technology of deep learning has developed rapidly. More and more new super resolution reconstruction algorithms have appeared in the software level, and achieved more practical effects than the traditional methods [1]. However, the existing deep learning super resolution reconstruction methods also have some problems, such as inaccurate restoration of image brightness space, image detail texture distortion or excessive sharpening [2]. The model proposed in this paper mainly uses the double regression thinking to form a closed-loop network by using the original regression network and the double regression tasks, and then connects each module in the network in series by using the jump connected direct channel, which can simplify the redundant construction of the model and improve the operation efficiency of the model.

II. RELEVANT THEORETICAL SUPPORT

A. U-Net network

U-Net network model is one of the most successful models in image segmentation, especially in medical image segmentation. This network model was put forward at the MICCAI Conference in 2015, and the number of references is still on the rise, and the theoretical performance of various models improved based on U-Net network model has been improved to a certain extent. The encoder and decoder structure adopted by U-Net network is a network model with different ideas from the classic GAN network. The model does not adopt the machine learning composition for training by the mutual game between generator and discriminator. However, the jump connection of its network model is a very classical design method, which greatly improves

DOI: 10.2478/ijanmc-2023-0055

the overall efficiency and performance of the network model. The U-Net network model is described in detail below [3].

U-net is a leap-forward model based on the full convolutional network, which adopts the complementary construction of encoder and decoder [4]. Because its network structure is in the shape of "U", it is called U-Net network. The U-Net network model is shown in Fig 1:

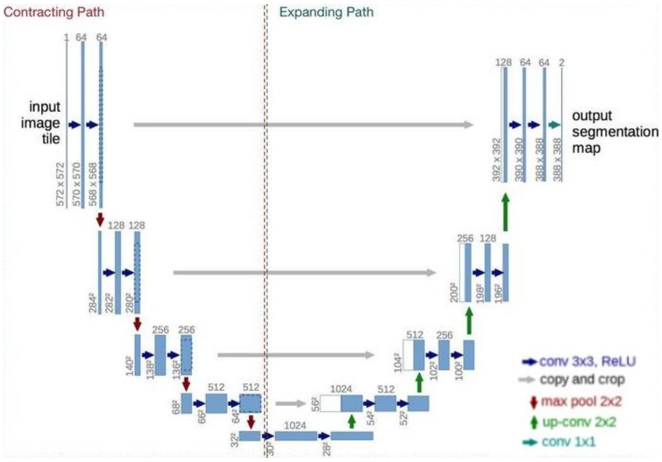

Figure 1. U-Net Network structure

The U-shaped structure of U-Net is shown in the figure. The network is a classical full convolutional network, which means that the network does not contain full connections. The input of U-Net network model enters from the left, and after two 3*3 convolution, the input image is cropped into a regular image with a resolution of 568*568. U-Net network is divided into encoder on the left and encoder on the right. In the middle, the convolution operations at both ends are layered and integrated by three skip connections.

The operation performed by the left encoder is a downsampling operation, which is composed of different convolution cores and maximum pooling layers. This part is called the compression path. The compression path consists of 4 different convolution modules and a 2*2 maximum pooling layer. Each module uses 3 effective convolution operations and 1 maximum pooling drop recovery respectively. After the downsampling operation of the above different modules, the relevant feature maps of the images are obtained. Then, the number of feature maps is doubled to obtain the feature maps with a size of 32*32.

The operation performed by the decoder on the right is an upsampling operation, consisting of a different deconvolution kernel with a 2*2 upsampling convolution layer. Each processing unit of the decoder is still composed of 4 different modules, which are 2 3*3* convolution modules, one upper sampling layer and one normal layer. The convolutional module multiplifies the size of the feature graph extracted previously by 2 through deconvolution operation, reduces its number by twice, and combines the feature graph obtained from the encoder. The size of the feature graph obtained at last is 388*388. Thus, the coding and decoding process of the whole network model is completed [5].

U-Net model not only connects the whole network in series by skip connection, but also increases the flexibility of the whole network, which greatly enhances the decoding efficiency. In addition to feature extraction of the image, deconvolution is used to restore the size of the image and promote each other before and after encoding and decoding, thus improving the running speed of the whole network model.

It can be seen that U-Net can not only ensure the global information of the image, but also consider the details of the image. At the same time, it can support a small amount of data training model. Based on the advantages of U-Net network, we choose to transform it and form the double regression lightweight network model in this paper.

B. Residual channel block RCAB

1) Residual block RB:

As CNN feature extraction network is widely used in deep learning, in scholars' general impression of CNN, the deeper the level of deep learning network, the stronger the expression ability of image features. Therefore, the research using CNN gradually expanded from Alexnet's 7-layer network structure to Googlenet's 22-layer network. However, with the deepening of the research, it is found that after CNN has reached a certain level, simply increasing the number of layers cannot achieve the expected classifier performance improvement. In addition, the convergence of the network also starts to slow down when the level continues to rise. During the experiment, it is also found that when the network level increases, the accuracy of network classification reaches saturation and even begins to decline [6].

ResNet came into being under such circumstances. Inspired by the concept of residuals commonly used in the field of computer vision, ResNet applied its concept to the construction of CNN model, and thus there was a basic structural block of residuals learning. ResNet maintains network complexity by balancing the size and number of feature graphs. Different from the general CNN, ResNet network is designed to learn residuals from the input image to the output image through a hierarchy with parameters [7]. ResNet residuals are learned by adding a short circuit between each layer. ResNet uses residual learning block to solve the degradation problem of deep network, so that CNN can train deeper network through this method. The residual structure is shown in the Fig 2.

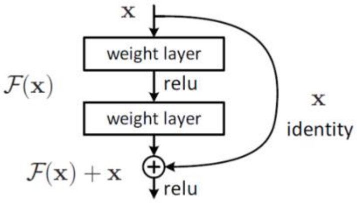
x
weight layer
relu
weight layer
$\mathcal{F}(\mathbf{x})$
x
identity
$\mathcal{F}(\mathbf{x}) + \mathbf{x}$
relu

Figure 2. Residual learning unit RB

2) *Channel attention mechanisms (CA)*

Attention mechanism is a special module structure in the deep learning network model. When processing some features of the image, the network model will give some pixels corresponding "special attention" to highlight some important features, ignore the irrelevant part, and pay more attention to the key information, which will improve the task quality and work efficiency to some extent[8].

At present, attention mechanism is more or less added to network models to improve the precision of image detail processing, especially in the image high-frequency texture and edge transition. Most of the existing attention mechanisms are excellent at deep learning. The best ones are the Squeeze-and-Excitation(SE), BAM and CBAM. The performance Excitation is different from those of other attention mechanisms. Taking the SE module as an example, the flexibility of the SE module lies in that it can quickly adapt to the operation mechanism of the existing network. It compresses the features of the image through the spatial dimension, and turns each two-dimensional feature channel into a specific real number [9]. The real number has a global receptive field at this time, and its output size also ADAPTS to the feature channel of the image when it is input. This can be very useful in many scenarios.

However, the self-attention mechanism also has its own disadvantages. SE only considers the internal channel information and ignores the importance of location information, while the spatial structure of the target in vision is very important. BAM and CBAM try to introduce location information through global pooling on channels, but this approach captures only local information, not long-scoped dependent information [10]. The attention mechanism used in this article uses the channel attention mechanism (CA) in conjunction with residual blocks. The CA mechanism captures not only cross-channel information, but also direction-aware and position-sensitive information, which enables the model to locate and identify the target area more accurately. This approach is flexible and lightweight, and can be easily plugged into existing classic mobile networks. The CA structure is shown in the Fig 3:

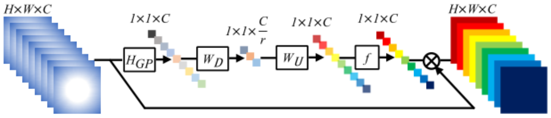

Figure 3. Channel attention (CA)

If the main part of the network model that connects residual blocks and long jumps focuses on the more informative components of LR features, is it feasible for the channel attention mechanism to extract the statistics between channels and further enhance the discrimination ability of the network? We integrate CA into RB and get residual channel attention block (RCAB). The structure is shown in the Fig 4.

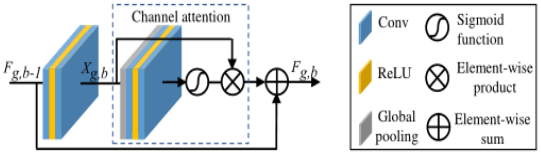

Figure 4. Residual channel attention block (RCAB)

III. Research content of this paper

A. Theoretical basis of double regression model

At present, there are few available LR-HR paired data in the known datasets, and in most cases, the HR images need to be down-sampled to obtain the corresponding LR images before model training. However, paired LR-HR data may not be available in real-world applications, and the underlying degradation method is usually unknown. For this more general case, existing SR models tend to produce adaptive problems and produce poor performance. In this paper, a dual regression model is constructed to adapt the SR model to the new LR data using both real-world LR data and paired synthetic data, introducing additional constraints on the LR data to reduce the space of possible functions [5].

$$I_{xLR} = d(I_{yHR}, \partial) \tag{1}$$

$$g(I_{xLR}, \delta) = d^{-1}(I_{xLR}) = I_{yE} \approx I_{yHR} \tag{2}$$

$$d(I_{yHR}, \partial) = (I_{yHR}) \downarrow_{S_f}, \{s\} \subseteq \partial \tag{3}$$

Where $I_{yHR} \otimes \kappa$ represents HR image, I_{yHR} convolution with fuzzy kernel κ, and n_σ additive σ Gaussian white noise with standard deviation. The degradation function defined in formula (4) is closer to the actual function because it takes into account more parameters than a simple down-sampled degradation function.

$$d(I_{yHR}, \partial) = (I_{yHR} \otimes \kappa) \downarrow_{S_f} + n_\sigma, \{\kappa, s, \sigma\} \subseteq \partial \tag{4}$$

Where $L(I_{yE}, I_{yHR})$ is the loss function between the output HR image after SR and the actual HR image, and $\psi(\phi)$ is the regularization term. The loss function most commonly used in SR is based on the pixel mean square error, also known as pixel loss.

Algorithm 1: Adaptation Algorithm on Unpaired Data.

```
Input: Unpaired real-world data: Su;
    Paired synthetic data: Sp ;
    Batch sizes for Su and Sp : x and y;
    Indicator function: Sm.
    1.Load the pretrained models P and u;
    2.while not convergent do
    3.Sample unlabeled data {xi} from SU ;
    4.Sample labeled data {(xi , yi)} from SP ;
    5.// Update the primal mode
    6.Update P by minimizing the objective:
    7. $\sum_{i=1}^{m+n} I_{S_p}(x_i)\iota_p(P(x_i), y_i) + \lambda\iota_D(D(P(x_i)), x_i)$
    8.// Update the dual model
    9.Update D by minimizing the objective:
    10. $\sum_{i=1}^{m+n} \lambda\iota_p(D(P(x_i)), x_i)$
    11.END
```

B. Loss function selection

In the training of network model, the generation counter minimum is calculated. Since the first half of the formula has nothing to do with the generator, the second half of formula (6) is taken in the actual training, and the T value is set to 1.

$$L_{Adversarial} = \min E_{I^{LR} \sim p_G(I^{LR})}\left[L_M(D_{\theta_G}(G_{\theta_G}(I^{LR})), T=1)\right] \tag{5}$$

$$L_{Vecoter}^{SR} = \sum_{i}^{m}(V_i(I^{HR}) \cdot V_i(G_{\theta_G}(I^{LR})) - \left\|V_i(I^{HR})\right\|^2)^2 \tag{6}$$

In addition to counter loss, in order to improve the texture detail of the generated image, this paper proposes a vector inner product loss function $L_{Vecoter}^{SR}$. Where V represents the vector before the loss function of the product inside the capsule is taken by the compression rectification function, that is, a 16-dimensional vector taken from the normalized layer of the network. The subscript of V i represents the number of classified sequences, and m is the total number of classified sequences. In the experiment, it is a binary classification of true and false. The value of i is 0 or 1.

C. Double regression nonlinear network model

We construct the network model based on U-Net network design. Our network model consists of two parts: original network and double regression network. We will introduce the details of the network in the Fig 5.

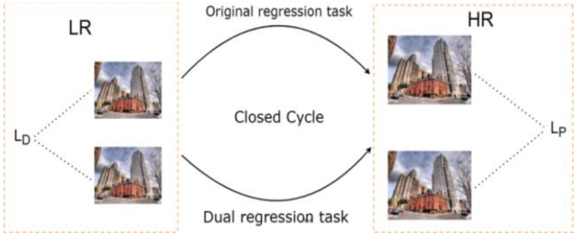

Figure 5. Double regression theoretical model

The network model follows the "U-shaped" design pattern of U-Net. Our double regression scheme connects the original regression structure with the double regression structure through the direct connection channel. In the original regression scheme, there is only one down-sampling module and one up-sampling module for simple full connection operation of images, which is suitable for smooth images with small transition of edge details, such as simple figure pictures under solid color background or simple uniform font pictures, etc. Of course, such pictures are relatively few in real life. Then, it is necessary to take into account the position information generated by the high frequency conversion of the image. If an image is divided into two equal pieces by any straight line and the high frequency conversion occurs in one of the two areas, it can be activated by a convolution operation and then output into the corresponding high resolution image through the residual-channel block and upsampling. If the position of the high-frequency conversion is uneven and irregular, it can be subsampled again to continue feature extraction, and the edge detail texture of the image can be processed to the extreme to form a corresponding closed-loop, so as to achieve the effect of super resolution reconstruction. Therefore, under the mutual restriction of the original network and double regression, the double regression network model can get a high-resolution image closer to the real environment. The specific network model is shown in Fig 6.

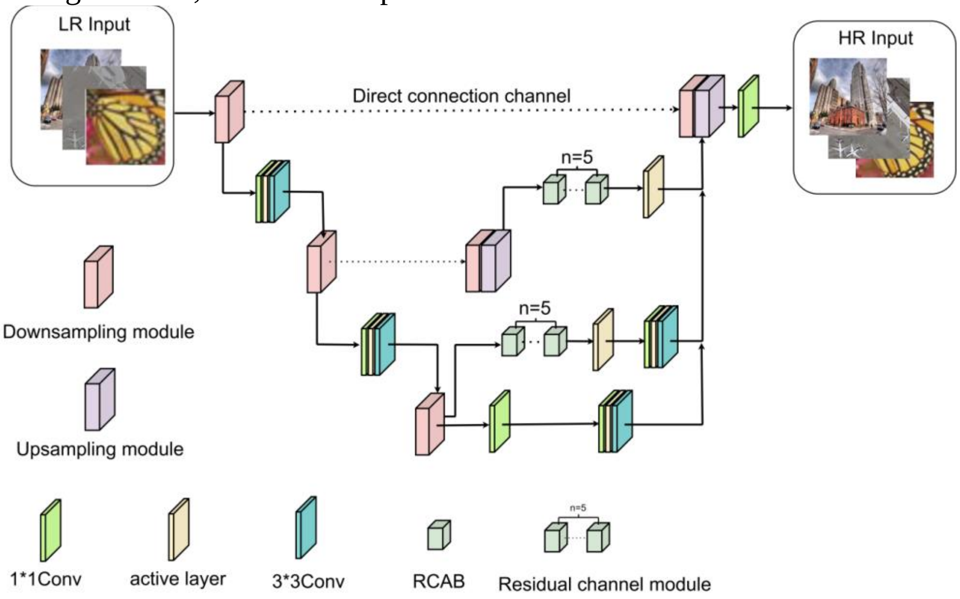

Figure 6. A double regression network training model based on U-Net network transformation is presented

IV. Experiment and Result

A. Experimental environment and setting

In order to adapt to the training environment of network model, the operating system is Microsoft Window 64-bit, CPU is E5 2698v3, memory is DDR4 128G, frequency is 3200MHZ, GPU is NVIDIA Titan V*3. The CUDA Version is 11.3. This experiment was run on the underlying environment of Anaconda with PyCharm as the compiler and an external third-party library composed of torch1.8.1, numpy1.23.5, visdom0.2.4, pytorch 1.9.0, etc. All networks share the same training Settings. We set the batch size to 18 to speed up the training process of the network model. We used the Adam optimizer to update the network parameters and set the initial learning rate at 10-4 and the training period at 200 epochs. When 30, 50 and 80 epochs were reached, the learning rate was multiplied by 0.2. The real-time loss function diagram drawn by the network is drawn by connecting MATLAB to the database.

This experiment adopts NTIRE2018 DIV2K data set to optimize the training model, which is specially used for the super resolution field of images. There are a total of 1000 2K resolution images, including 800 high-resolution images for training, 100 verification images and 100 test images. For part of the training set, rotation, scaling, translation and other methods were used to enrich the data set for data enhancement and more adequate training model. In addition, the 800 2K resolution images were downsampled by using the Bicubic method to obtain the corresponding 800 low-resolution images to enrich the training set. The test set used is a public benchmark data set widely recognized in the super resolution field, including Set5, Set14, and BSD100.

Figure 7 shows the comparison between our network model and the two classic network models. Compared with SRGAN, our edge texture is finer and the local brightness is closer to the original HD image. Figure 8 shows a graph of our network model and the relative PNSR value of SRGAN.It can be seen that our network model occupies certain advantages based on the digital model as the measurement standard and the premise of a large data set.

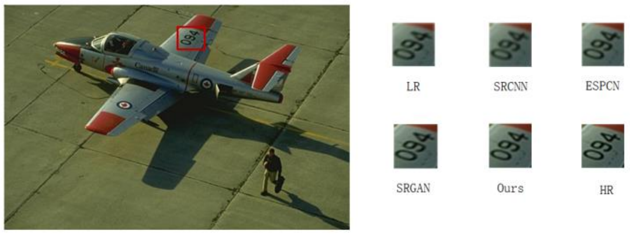

Figure 7. Sample graph of data set

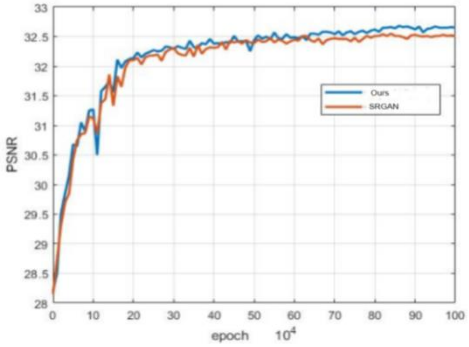

Figure 8. PSNR data comparison graph

In the experiment, PSNR is used as the key evaluation index of image. The PSNR training results of the network model based on Set5 data set are shown in the figure. As can be seen from the figure, the bi regressive linear network model has a relatively stable convergence process, and can stably generate high-quality super-resolution images after a certain number of training times.

B. *Experiment settings*

In the experiment, bicubic interpolation is used to downsample the original high resolution color image and obtain the corresponding low resolution color image. For training with 800 high-resolution images, the model proposed in this paper was used. All images were pre-trained by subtracting the average RGB value of DIV2K data set, and low-resolution images in DIV2K training set were cut into 48*48 image blocks, 16 color image blocks were used for each batch as input. This chapter uses the Adam optimizer with the parameter,,. The initial learning rate is halved every 500 cycles. In the attention mechanism, set r=16 and the convolution kernel to size 1*1. The other convolution kernels are of size 3*3. The boundary of each feature graph is zero-filled to ensure that its space size is the same as the input size after the convolution operation. The number of filters is 64. The performance of the model is evaluated by comparing the peak signal-to-noise ratio and structural similarity with other classical super resolution models.

C. *Ablation experiment*

In order to analyze the roles and contributions of different modules in the experimental model, this section proves the effectiveness of the modules proposed in the theoretical model through ablation experiments. As shown in the table, all the networks in the ablation experiments in this section had the same network depth, the up-sampling factor was 2, and the training set containing 800 images and Set5 were used as the test set. "Yes" in the table indicates that the network retains the structure, "No" indicates that the network deletes the structure.

As can be seen from the table, PSNR decreases by 0.128dB when only residual blocks are removed. When only the channel attention mechanism was removed, PSNR decreased by 0.134 db. When the residual block and the channel attention mechanism are removed together, the PSNR decreases by 0.24dB. It can be seen that the double regression model proposed in this paper significantly improves the performance of super resolution reconstruction by using the residual channel attention block, especially when the residual block and the channel attention mechanism are combined. The PSNR of reconstructed images directly increased by 0.24 dB compared with that without using the attention mechanism. When the direct channel was removed, the PSNR decreased by 0.059 dB; PSNR was reduced by 0.037 dB when the activation function was modified to use ReLu. Therefore, from the above analysis, it can be concluded that the residual channel block, direct path and the use of PReLu activation function proposed in this paper can improve the performance of image super resolution. The specific data are shown in Table 1.

TABLE I. ABLATION EXPERIMENT

PSNR	CA	RB	Direct connection path	PReLu
37.850	No	Yes	Yes	Yes
37.844	Yes	No	Yes	Yes
37.738	No	No	Yes	Yes
37.919	Yes	Yes	No	Yes
37.941	Yes	Yes	Yes	No
37.978	Yes	Yes	Yes	Yes

D. Comparison of experimental results

In order to verify the feasibility of optimization and improvement in this chapter on the original network, the classical Bicubic interpolation (SRCNN), DRCN, ESPCN and SRGAN among the traditional super resolution algorithms will be selected for comparative experiments. Meanwhile, the improved module used in this chapter will be used for ablation experiments. The test set used in this chapter is Set5, Set14 and BSD100 samples. Since the original size of the images is too large for display, the experiment will scale the displayed images and cut part of them according to the original size, so as to compare the detailed effects of the reconstructed images by different algorithms. The specific data are shown in Table 2.

TABLE II. COMPARISON OF ALGORITHMS FOR DIFFERENT DATA SETS

Method	Set5 PSNR/SSIM	Set14 PSNR/SSIM	BSD100 PSNR/SSIM
Bicubic	32.40/0.9589	31.32/0.9521	32.87/0.9563
SRCNN	33.36/0.9460	33.78/0.9366	33.57/0.9423
DRCN	33.57/0.9432	33.99/0.9419	33.66/0.9410
ESPCN	34.12/0.9439	34.26/0.9412	34.23/0.9356
SRGAN	34.26/0.9356	34.89/0.9256	34.56/0.9246
Ours	34.12/0.9326	34.56/0.9247	34.57/0.9232

The following figure shows the reconstructed results of each algorithm in the Set14 dataset. Direct use of Bicubic interpolation (Bicubic) image can be seen to be significantly fuzzy, image quality is poor, all algorithms compared with bicubic interpolation (Bicubic), image quality is improved. SRCNN uses three convolutional layers to reconstruct the image, which is not clear in terms of image details, which is also caused by insufficient depth of network layers and insufficient learning of image features. While the SRGAN network is deep enough, the overall detail of the image reconstructed using the generated counter network is perfect. Ours in this paper is oriented towards PSNR index, compared with other methods that use to generate antagonistic network. The specific comparison is shown in Fig 9.

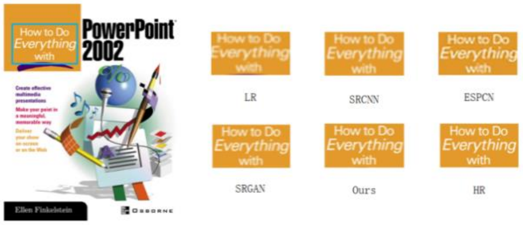

Figure 9. Comparison of ppt details

As can be seen from the hat detail in the image below, a high PSNR value does not necessarily help reconstruct the image detail. Ours method combines the anti-loss function and perceived loss. Based on the PSNR value-oriented generation network as the pre-training model, the overall details of the reconstructed image are clear and the image quality is better. The specific comparison is shown in Fig 10.

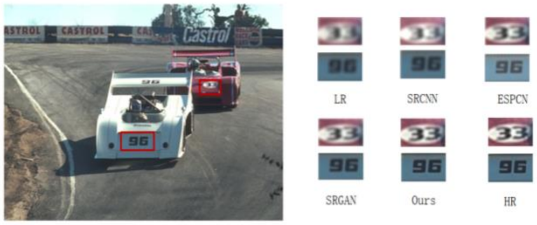

Figure 10. Comparison of baby details

V. Conclusions

A double nonlinear regression scheme is proposed for paired and unpaired data. On paired data, we introduce an additional constraint by reconstructing LR images to reduce the space of possible functions. Therefore, we can significantly improve the performance of the SR model. In addition, we focused on unpaired data and applied a dual regression scheme to real-world data. We performed ablation studies on the bi regression protocol, and models using the bi regression protocol showed better performance across all data sets compared to baseline. These results suggest that the dual regression scheme can improve HR image reconstruction by introducing additional constraints to reduce the space of the mapping function. We also evaluated the impact of our double regression scheme on other models. Compared with other classical algorithms, we compared PSNR results, SSIM results and visual data set images from two levels of double magnification and quadruple magnification, and it can be seen that the improved algorithm is significantly superior to the classical algorithm. Since there are many scenarios with super resolution, we only choose the ones that are biased towards buildings, trees, digital and other related directions in terms of data set. For the algorithm adaptation in medical and other professional fields, there may be some missing problems, which need to be further improved in the follow-up research.

References

[1] Nie L, Lin C, Liao K, et al. A view-free image stitching network based on global homography – Science Direct[J]. Journal of Visual Communication and Image Representation, 2020, 73.

[2] Zhang J, Wang C, Liu S, et al. Content-Aware Unsupervised Deep Homography Estimation [M]. Springer, Cham, 2020.

[3] Detone D, Malisiewicz T, Rabinovich A. Deep Image Homography Estimation [J]. 2016.

[4] Japkowicz N, Nowruzi F E, Laganiere R. Homography Estimation From Image Pairs With Hierarchical Convolutional Networks[C]// The IEEE International Conference on Computer Vision (ICCV). IEEE Computer Society, 2017.

[5] Guo H, Liu S, He T, et al. Joint Video Stitching and Stabilization from Moving Cameras [J]. IEEE Transactions on Image Processing, 2016.

[6] Le H, Liu F, Zhang S, et al. Deep Homography Estimation for Dynamic Scenes [C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020.

[7] Watson J, Hughes J, Iida F. Real-World, Real-Time Robotic Grasping with Convolutional Neural Networks[C]// Conference Towards Autonomous Robotic Systems. 2017. Zhaobenben, yinxudong, Wang Wei GitHub data crawler based on scrapy [J] Electronic technology and software engineering, 2016 (06): 199-202.

[8] Nguyen T, Chen S W, Shivakumar S S, et al. Unsupervised Deep Homography: A Fast and Robust Homography Estimation Model [J]. 2017.

[9] Nie L, Lin C, Liao K, et al. Unsupervised Deep Image Stitching: Reconstructing Stitched Features to Images [J]. 2021.

[10] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition [J]. Computer Science, 2014.