修 士 論 文 概 要 書

Master's Thesis Summary

Date of submission: 01/23/2023 (MM/DD/YYYY)

専攻名（専門分野） Department	Computer Science and Communications Engineering	氏 名 Name	Yun Liu	指 導 教 員 Advisor	Hiroshi Watanabe 印 Seal
研究指導名 Research guidance	Audiovisual Information Processing	学籍番号 Student ID number	CD 5120FG45-1		
研究題目 Title	A Prior-Guided Face Image Super-Resolution Network Based on Attention Mechanism				

1. Introduction

Due to the limitation of camera sensor hardware or the distance between the camera and the target, most of the captured face images are of poor quality. By using face image super-resolution, we can obtain high-resolution face images, thus providing help for subsequent tasks like face detection, face recognition.

Compared with general images, face images have distinctive features (e.g., the positions of facial organs are approximately the same), which can be applied in super-resolution tasks.

This paper proposes 2 approaches for face image super-resolution. In the method 1, we present a CNN-based approach in which we use attention mechanism as well as face prior information with the purpose of producing high-quality face images. The whole network contains two independent branches, one for predicting face parsing maps and the other for coarse super-resolution. We use a fusion module to integrate the outputs of these two branches to obtain the final super-resolution image. In the method 2, we extend the method 1 by using GAN to generate richer details of faces.

2. Related technology

2.1 SRResNet [1]

SRResNet is a CNN-based super-resolution approach. The basic module of SRResNet is called the residual block, in one residual block there are two convolutional layers, two BN layers, an activation function, and a skip connection inspired by ResNet. There are B residual blocks in this model. Pixelshuffle layers are utilized to reconstruct the final SR image.

2.2 Attention mechanism

The human attention mechanism enables humans to quickly sift through the vast amount of information to find useful information. Attention mechanism in deep learning mimics human's attentional thinking and has been widely applied in tasks such as natural language processing and computer vision with remarkable results. Channel attention is an important part of the attention mechanism, SENet [2], a model based on channel attention, whose key idea is to model the importance of individual feature channels and then enhance the important channel information and suppress the unimportant ones.

3. Proposed methods

3.1 Proposed method 1

The whole framework of method 1 is shown in Fig. 1 and the basic block of its SR branch is shown in Fig. 2. In the basic block, we use convolution kernels of different sizes in the feature extraction part and fuse the features extracted from each convolution layer to improve the feature extraction capability of the model. A softmax operator is used to calculate the weights of each channel and these weights are multiplied with the feature map to emphasize the useful channel features and suppress the useless ones. In the prior branch, the hourglass structure is used to predict the parsing map of the face image. Low resolution face images provide limited information, it will be helpful to reconstruct the facial components if the accurate parsing map can be predicted and used as the prior information.

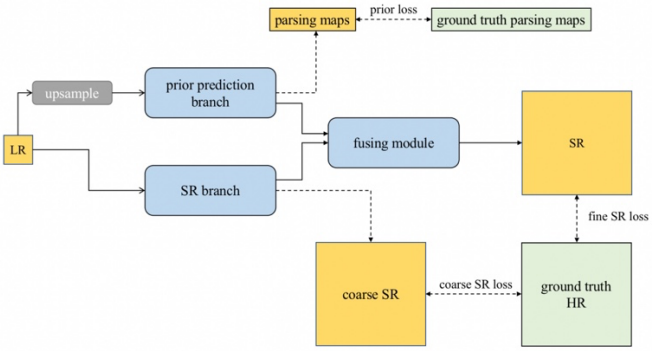

Fig. 1. Overall framework of the proposed method 1.

The Fusing module is used to fuse the information from the two branches for the final information integration and reconstruction. It is worth mentioning that the super-resolution branch and the prior prediction branch are trained simultaneously and the whole network is end-to-end.

The loss function we use consists of three components: coarse SR loss, prior loss, fine SR loss.

coarse SR loss: MSE loss between coarse SR image and ground truth HR image.

prior loss: MSE loss between predicted parsing maps and true parsing maps.

fine SR loss: MSE loss between final SR image and ground truth HR image.

Total loss:

$L_{content} = coarse\ SR\ loss + prior\ loss + fine\ SR\ loss$

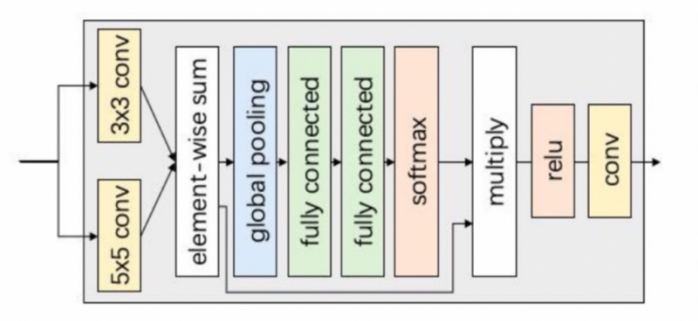

Fig. 2. Basic block of SR branch.

3.2 Proposed method 2

We use GAN in this proposal, the generator directly adopts the model proposed in method 1. And as for the discriminator, we use the discriminator in SRGAN [1], which contains multiple convolutional layers. The optimization of GAN is a "min-max" process, and for our face super-resolution task this problem converts to:

$$\min_{\theta_G}\max_{\theta_D} E_{I^{HR}\sim p_{train}(I^{HR})}[logD_{\theta_G}(I^{HR})] + E_{I^{LR}\sim p_G(I^{LR})}[\log(1 - D_{\theta_D}(G_{\theta_G}(I^{LR})))]$$

4. Experiments and results

4.1 Dataset

We use 28,000 images from CelebAMask-HQ Dataset [3] to train the model and 1000 images are used for testing. This dataset contains face images and their corresponding 11 face parsing maps, including jawline, eyebrows, eyes, nose, mouth and ears. The size of the HR face image is 128×128, and we down-scale the HR image by the factor of 8 to obtain the LR face image, whose size is 16×16.

4.2 Quantitative evaluation

The results of Bicubic, SRResNet, proposed method 1, proposed method 2 and Ground truth are shown in Table 1, we can notice that the PSNR and SSIM of Bicubic are the lowest, and our proposed method 1 outperforms SRResNet in both PSNR and SSIM, with an increase of 0.19 dB and 0.0029 in the two metrics, respectively. Compared with SRResNet and proposed method 1, proposed method 2 has lower PSNR and SSIM, but on the other hand has the lowest LPIPS value among all methods, which means better visual perceptual quality.

Table 1. Experimental results (8×super-resolution).

Method	PSNR(dB)	SSIM	LPIPS
Bicubic	20.72	0.5200	0.5351
SRResNet	24.47	0.7201	0.1854
Proposed method 1	**24.66**	**0.7230**	0.1900
Proposed method 2	23.38	0.6689	**0.1018**

4.3 Qualitative evaluation

To show the performance of the proposed methods, we compare the qualitative results with SRResNet and Bicubic. As shown in the second row of Fig. 3, the human eye area obtained by the SRResNet method is blurred, and even the eyes seem to disappear. In addition, it can be seen from the first row that SRResNet is weak in reconstructing the tooth parts and cannot accurately reconstruct the shape of the teeth. The proposed method 1, on the other hand, recovers the facial contours as well as the mouth and eye edges better. The proposed method 2 also has some improvement for the recovery of facial details. From the test images, we can see that both proposed methods have improved the performance.

Fig. 3. Qualitative evaluation result (image source: [3]).

5. Conclusion

We propose two new methods for face image super-resolution. The proposed method 1 is a CNN-based model, we add a multi-scale channel attention mechanism and face prior information based on SRResNet. Our model contains three main parts, namely SR branch, prior prediction branch and fusing module. These three parts are optimized by their respective loss functions. The proposed method 2 is an extension of proposed method 1 and incorporates the training idea of GAN. Experiments reveal that the proposed method 1 has improved PSNR and SSIM values compared with SRResNet and Bicubic; although the PSNR and SSIM of our second proposal have decreased, its LPIPS is optimal, which means better subjective perception results for human eyes.

6. Reference

[1] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, "Photo-realistic single image super-resolution using a generative adversarial network," in CVPR, pp. 105-114, Jul. 2017.

[2] J. Hu, L. Shen, G. Sun, "Squeeze-and-excitation networks," in CVPR, pp. 7132-7141, Jun. 2018.

[3] http://mmlab.ie.cuhk.edu.hk/projects/CelebA/CelebAMask_HQ.html). Under the agreement that The CelebAMask-HQ dataset is available for non-commercial research purposes only.

A Prior-Guided Face Image Super-Resolution Network Based on Attention Mechanism

A Thesis Submitted to the Department of Computer Science and Communications Engineering, the Graduate School of Fundamental Science and Engineering of Waseda University in Partial Fulfillment of the Requirements for the Degree of Master of Engineering

January 23rd, 2023

Yun LIU

(5120FG45-1)

Advisor: Prof. Hiroshi Watanabe

Research guidance: Research on Audiovisual Information Processing

Acknowledgements

First of all, I would like to express my heartfelt gratitude to Professor Hiroshi Watanabe for providing an excellent environment for this research and for his careful guidance in my daily studies. In addition to his academic guidance, he also warmed my heart by providing encouragement and support during those days when I was feeling pressured.

Secondly, I would also like to thank Mr. Takeshi Chujoh, Mr. Tomohiro Ikai, Mr. Takuya Suzuki, and Mr. Zheming Fan from Sharp Corporation for their guidance and assistance in our joint research, which provided me with many valuable suggestions and helped me tremendously in writing my thesis.

In addition, I would like to express my gratitude to our lab members, they gave me a lot of suggestions and comments in the seminars.

Finally, I would like to express my gratitude to my parents for supporting my dream of studying in Japan, giving me unreserved support both economically and spiritually, and giving me all their love.

Abstract

Low resolution(LR) face images can only provide very little information for people, which poses a challenge for subsequent face detection, recognition or segmentation techniques. The face image super-resolution technique can obtain a high resolution face image by processing one or several low resolution images, thus providing a good basis for subsequent processing.

We propose two face image super-resolution methods in this thesis, aiming to reconstruct high quality face images from a low resolution input. The proposed method 1 introduces an attentional multi-scale feature fusion block, which aims to improve the representation capability of the neural network by emphasizing the important feature maps and suppressing the unimportant ones. In addition, the facial prior information is utilized by adding a separate prior branch, an hourglass structure is used. The proposed method 2 adds GAN to the proposed method 1, aiming to generate more face details. Experiments demonstrate that the face images generated by our proposed methods exhibit noticeable quality improvement compared to the low resolution images and other SR approaches.

Keywords: Face image super-resolution, channel attention, face prior, generative adversarial networks

List of contents

List of figures

List of tables

Chapter 1 Introduction

1.1 Research background

In recent years, the security of public places has been an issue of great concern, especially in densely populated places. Surveillance systems are becoming increasingly popular for urban management and security control. However, due to factors such as the installation location of the surveillance equipment and the inadequate hardware conditions of the equipment itself, the face images produced are sometimes blurred, mutilated, noisy, small in size or other problems. In addition to the problems with the surveillance equipment itself, the compression process taken during the transmission and storage can also degrade the quality of the face images. Degraded face images may lack critical details and will adversely affect subsequent face-related tasks.

These problems can be solved by two types of solutions: hardware-based approaches and software-based approaches. Hardware mainly refers to image sensors. The spatial resolution of images captured with a certain device (e.g., camera or smartphone) is usually limited by the performance of the vision sensors. Increasing the number of sensors can lead to higher image resolution, but the corresponding hardware also requires higher costs. In order to achieve a balance between hardware budget and performance, a number of software-based approaches have been proposed that can save costs, eliminate dependence on expensive hardware devices, and achieve more satisfying results. Image super-resolution is a technique which uses specific algorithms to process a low resolution image to obtain a corresponding high resolution image. Compared to low resolution images, high resolution images tend to have clearer details and provide a better visual experience.

The human face has structured features, so face image super-resolution is a specific kind of image super-resolution task. High quality face images are very useful for face detection and recognition, but in some scenarios, the resolution of the obtained images is not high and needs to be repaired to be used effectively. However,

when the scale factor is too large (e.g. eight times), most of the reconstruction performance will be drastically degraded. How to use existing super-resolution techniques to deal with very low resolution face images is a challenge in face super-resolution and there are many researchers working on this topic.

1.2 Single image super-resolution

Classified by research methods, super-resolution algorithms can be divided into three main mainstream methods, namely interpolation-based, reconstruction-based and learning-based methods.

The interpolation-based method is based on the similarity between neighbouring pixels, and the current pixel is inferred from the neighbouring known pixels. This approach does not rely on learning huge amounts of data, so it is straightforward and fast to implement. Common interpolation-based super-resolution methods include: Nearest, Bilinear, and Bicubic interpolation.

The second method is based on reconstruction. The Projection onto Convex Set was first proposed by Stark and Oskoui [1], which takes a LR sequence as the processing object and makes optimal use of the prior knowledge to generate a high resolution image, the performance is enhanced compared to interpolation methods. Irani et al. [2] proposed Iterative Back-Projection method, the key idea is to project the errors of the simulated low-resolution image and the real low-resolution image onto the estimated high-resolution image, and keep repeating this operation so that the high-resolution image gradually approaches real high-resolution image. Schultz et al. combined the ideas of Probability Theory with the task of super-resolution and worked out a super-resolution method based on the maximum posterior probability(MAP).

Learning-based methods are gradually becoming dominant in recent years. Such methods aim at learning the mapping between high resolution and low resolution images; Yang et al. [3] use sparse representation for image super-resolution. Chang proposes a neighbour embedding method, which utilize the relationship of image blocks in low resolution space to constrain the relationship of image blocks in high

resolution space, obtaining better reconstruction quality. With the fast development of AI, a series super-resolution approaches have emerged. Dong et al. proposed SRCNN [4] using convolutional neural networks, which adopts a three-layer CNN structure. The authors of SRCNN have further proposed FSRCNN [5], which reduces the computational effort and speeds up the convergence of the network compared to SRCNN. In the VDSR [6] published in 2016, the authors deepened the number of layers of the network to 20 and used a larger perceptual field, adding the concept of residual learning to achieve a better performance. ESPCN [7] reconstructs images by sub-pixel convolution, placing the up-sampling module at the last layer of the network, all feature extraction is conducted in low resolution space, which greatly reduces the computational effort. Structures such as SAN [8], RCAN [9] and RDN [10] proposed in recent years have also achieved impressive results in image super-resolution.

1.3 Research Objectives

The purpose of this paper is to optimize the existing super-resolution algorithm and make it applicable for face images.

The main idea is to incorporate the latest feature extraction method, the attention mechanism, to extract richer information about face features. In addition, considering the special characteristics of faces, our approach incorporates facial priors as auxiliary information. Generative adversarial networks, with their pioneering design concepts and stunning generative capabilities, have been broadly utilized in diverse tasks in computer vision. In the second proposal of this thesis, we also try to adopt GAN and achieve satisfactory results.

This paper is dedicated to using face image super-resolution to generate high-quality face images, we have 4 objectives:

1. Explore the classic image super-resolution algorithm SRResNet [11] for potential improvement points.
2. Introduce the channel attention mechanism as well as prior information to the

classical approach to compare the performances of various schemes.

3. Add the idea of GAN for model training and compare it with CNN-based methods to observe the generative results.

4. Analyse the results and explore new application scenarios.

1.4 Outline of thesis

The structure of the thesis is summarized as follows:

Chapter 1: We describe the causes of low resolution face images and benefits that super-resolution techniques can bring. Next, the development history and categories of image super-resolution are described in detail. Finally, the research objectives and the outline of this paper are presented.

Chapter 2: We present some related works, ranging from the principles underlying CNN and GAN to super-resolution methods using these two techniques. We also present some techniques and basic knowledge related to the proposed approaches in this paper, including attention mechanism and the content of face prior information.

Chapter 3: We present the two proposals separately. The network structure and principles of each method, and their loss functions are included. In the proposed method 1, we specify the structure and purpose of the following three modules: SR branch, prior prediction branch and fusing module. In the proposed method 2, we give details of the optimization process based on the GAN method.

Chapter 4: The experimental results are demonstrated. We trained and tested the different methods on a publicly accessible dataset. We evaluate different methods in both subjective and objective terms, and the strength of our proposals is demonstrated.

Chapter 5: We summarize the contents of this thesis.

Chapter 2 Related Technology

2.1 Convolutional Neural Networks

With its ingenious design and superior performance, convolutional neural networks (CNNs) are often utilized in various tasks where the object is an image, such as image classification, recognition, etc. Unlike normal neural networks, convolutional neural networks change the connections between neural layers to sparse connections. The advantage of this structure is that it brings a reduction in the amount of parameters and reduces the model complexity. Another feature of CNN is that the parameters are shared, with each local connection using the same parameters, each convolutional kernel extracting one feature and multiple kernels extracting multiple features. In the coming paragraphs we will specify some important components of CNN.

2.1.1 Introduction of convolutional layer

The process of convolution deals with a block region of pixels in an image, rather than individual pixels, deepening the neural network's understanding of the image. A convolution operation is shown in Fig. 2. 1, where the 3×3 block in the middle is called convolution kernel. The information in the image is extracted by sliding the convolution kernel, and the output is called feature map. In addition to the kernel size, the convolution operation has two other important parameters: the stride and the padding value. The stride is the distance of each stroke of the kernel and is usually 1 or 2, depending on the situation.

The feature map gained by moving the kernel directly on the input image will be smaller in size than the original image, padding solves this problem by first zero-filling around the input image, which is equivalent to enlarging the original image, and then doing convolution on the it, which ensures that the dimensions of the output feature map are identical to those of the input.

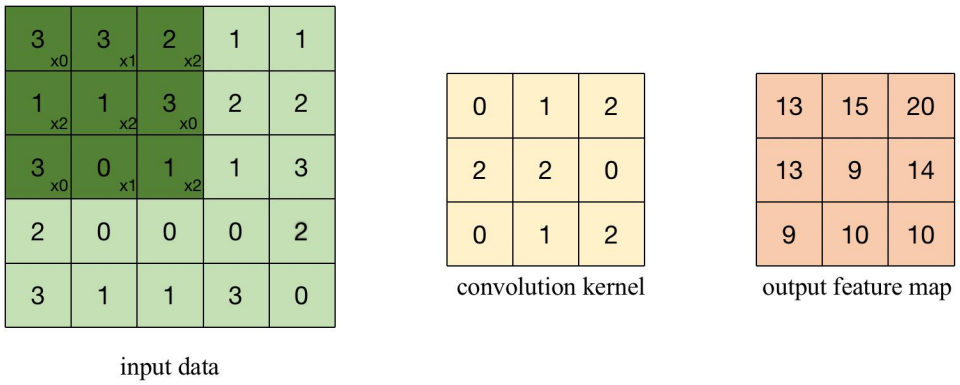

Fig. 2. 1. Convolution layer.

2.1.2 Introduction of pooling layer

Max pooling and average pooling are two typical pooling methods. The process of max pooling is shown in Fig. 2. 2 , for a image block which contains four pixels, the output of its max pooling is the biggest value in that image block, and this maximum value is used to represent the original block. In some tasks, the variation in image size does not affect the reliability of the results, such as image recognition, where a large image may be recognized by a CNN model in the same way as a down-sampled image. Pooling operations can compress image information while keeping image features largely unchanged, in order to remove redundant information and reduce model parameters.

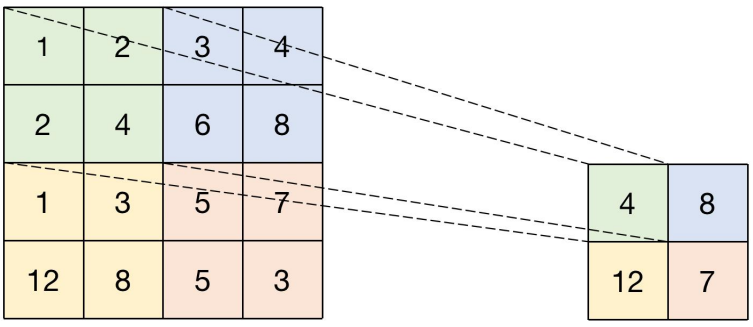

Fig. 2. 2. Max pooling layer.

2.1.3 Introduction of fully connected layer

A diagram of the fully connected layer structure is shown in Fig. 2. 3, which shows that in the hidden and output layers, each neuron receives information from all the neurons in the previous layer. However, there is an obvious disadvantage: the parameter number of fully connected layer is usually very large.

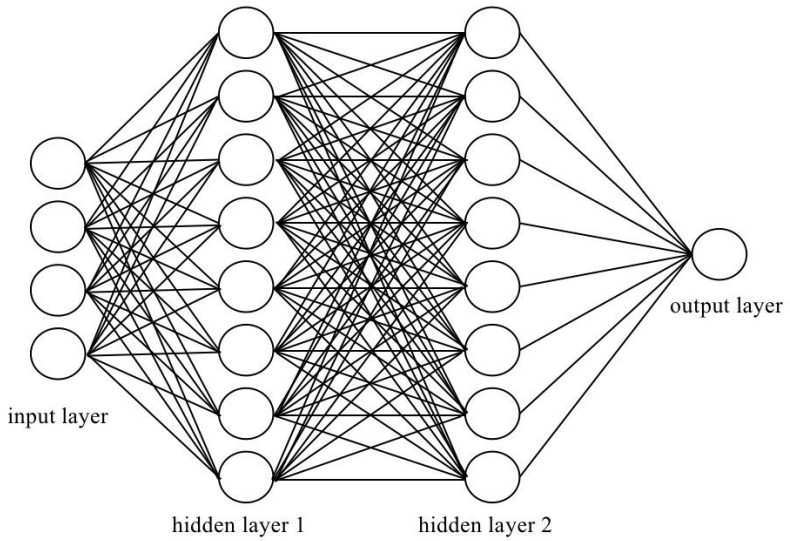

Fig. 2. 3. Fully connected layer.

2.2 Generative Adversarial Network

The basic structure of a Generative Adversarial Network (GAN) [12] is shown in Fig. 2. 4. The network is consisted of two components: a generator (G) and a discriminator (D). GANs have brought new ideas in data generation and are broadly used in a variety of tasks. Take image generation as an example, the input to G is usually random noise, after processing by the deep network structure, the output is an image, called the generated image. The input to D is a real image and a generated image, and its goal is to decide the authenticity of the current input. The training of GAN is a process in which the G and the D confront each other. More specifically, the G is trained to produce results closer to the real image through continuous iterations, while the D is trained to improve its ability to discriminate between generated fake images and real images. During training, the two networks play against each other and

their performance is gradually enhanced. When the D finally fails to classify the fake and real images, the neural network reaches a state of convergence and the parameters of the G are optimized.

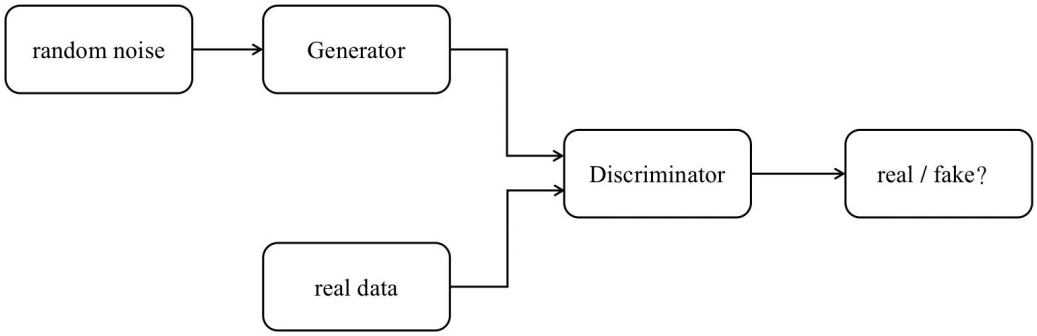

Fig. 2. 4. General structure of GAN.

The training of GAN is in fact a "min-max" process, where the objective function is defined by the following equation.

$$\min_G \max_D V(D,G) = E_{x\sim P_{data}(x)}[\log(D(x))] + E_{w\sim P_w(w)}[\log(1-D(G(w)))] \tag{2.1}$$

The above equations are essentially two optimization problems, where D and G are optimized respectively. When the D is trained, the optimization objective becomes

$$\max_D V(D,G) = E_{x\sim P_{data}(x)}[\log(D(x))] + E_{w\sim P_w(w)}[\log(1-D(G(w)))] \tag{2.2}$$

where x is from the real data, y is random noise and G(y) is the output of G. The optimization target of the D is that the output of the D is approaching 1 when the input is real data x, and conversely, the output of the D is approaching 0 when the input is a false sample.

In optimizing the G, the objective function becomes the following equation:

$$\min_G V(D,G) = E_{z\sim P_z(z)}[\log(1-D(G(z)))] \tag{2.3}$$

For the G, the goal is to generate data that is as realistic as possible to "fool" the D. In other words, the generator wants D(G(z)) to be as large as possible, and another

way to maximize this value is to minimize 1 - D(G(z)). The combination of the objective functions of the two networks is the 'min-max' problem in equation 2. 2.

GAN has been extensively applied in image generation domain due to its subtle structural design and excellent generation results.

2.3 Image super-resolution based on deep learning

Current deep learning-based image super-resolution algorithms typically use CNNs, with some works using GANs. This section will describe the general principles of CNN-based and GAN-based image super-resolution algorithms and typical approaches.

2.3.1 CNN-based methods

SRCNN is the first CNN-based image super-resolution approach, with a very simple structure containing only three convolutional layers. With the swift evolution of deep learning, more and more new techniques are being introduced to super-resolution tasks, achieving amazing results. A typical super-resolution network structure will consist of two steps: feature extraction and image reconstruction. The purpose of feature extraction part is to extracts deep features through convolutional operations, and the image reconstruction part integrates the features to obtain the super-resolution image. Because one of proposals is based on SRResNet, this subsection will use SRResNet as an example to describe the general structure of CNN-based super-resolution method, the structure is illustrated in Fig. 2. 5.

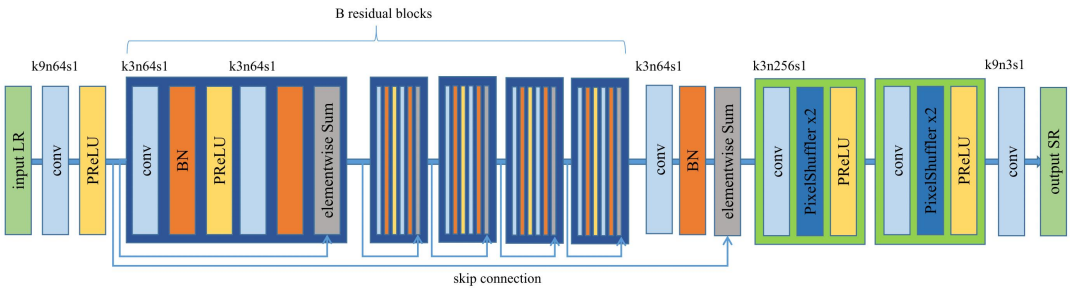

Fig. 2. 5. Structure of SRResNet.

The basic module for the feature extraction part of SRResNet is called residual

block, and we can see that in one residual block there are two convolutional layers, two BN layers, an activation function, and a skip connection inspired by ResNet [13]. There are B residual blocks in this model. The input to the image reconstruction part is the extracted deep image features, which are re-assembled by the PixelShuffle layer, which is shown in Fig. 2. 6. The purpose of PixelShuffle is to transform a feature map of size $W \times H \times r^2$ into a feature map of size $rW \times rH$.

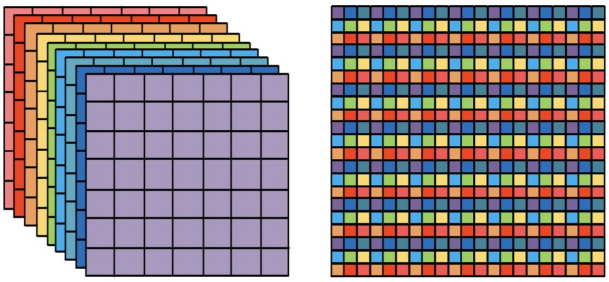

Fig. 2. 6. Schematic of Pixelshuffle layer.

2.3.2 GAN-based methods

Compared with the methods using CNN, GAN-based image super-resolution methods are usually better at generating some detailed texture information, and the image quality tends to better match human visual perception. The GAN-besed method is typified by SRGAN, whose generator structure is the SRResNet introduced above, and the discriminator is shown in Fig. 2.7.

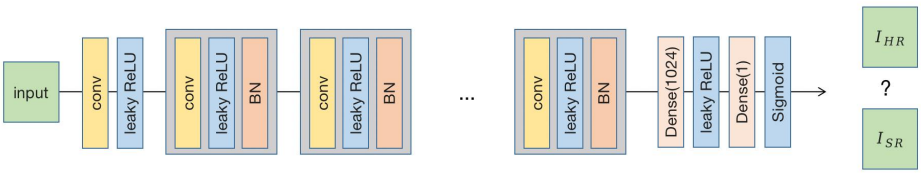

Fig. 2. 7. Discriminator in SRGAN.

The input to discriminator is the real HR image and SR image generated by the generator. The discriminator has eight convolution blocks, each with one convolution operation, one activation function, and one BN layer, and the stride is alternately set

to 1 and 2. The convolution with a stride of 2 works similarly to pooling, reducing the feature map size. The feature maps obtained after multiple convolution layers are then passed through two dense layers and together with a sigmoid module. The discriminator output is approximately 1 when the input is real, and closer to 0 when the input is judged to be fake.

ESRGAN [14] is a modified approach based on SRGAN, using Residual-in-Residual Dense Block (RDDB) rather than residual block in the network, which the authors believe can improve the model's capability and make it easier to train. In the discriminator part, the relativistic discriminator is used, aiming to make the discriminator learn "one image is more real than another" instead of "whether an image is real or fake".

2.4 Attention mechanism

The human attention mechanism helps humans to select more useful information from complex content. The attention mechanism in deep learning mimics the human brain and tries to extract the features that are more salient to the target task from the vast amount of features. Attention mechanisms are broadly adopted in natural language processing(NLP), and recently have also been applied to computer vision. Since the proposed approaches use the channel attention, the basic concept of channel attention will be highlighted in this section.

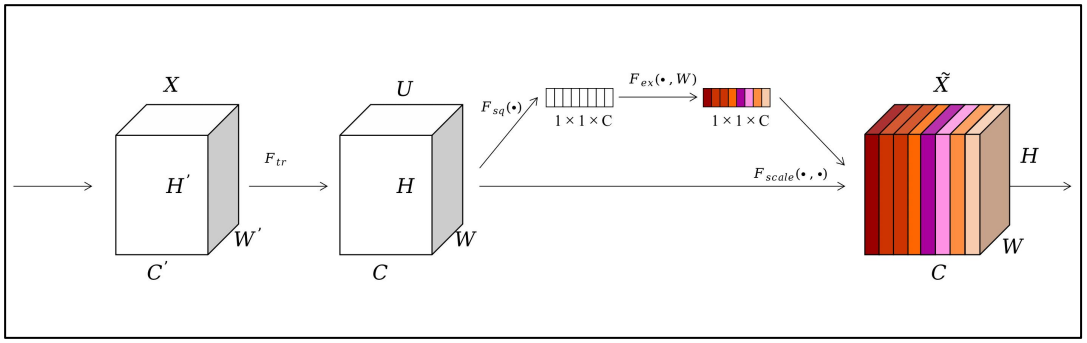

Fig. 2. 8. Channel attention mechanism.

SENet [15] proposes the concept of channel attention, the key idea of which is to model the importance of each channel, and then enhance the significant channel

information and suppress the unimportant ones, the diagram of channel attention is shown in Fig. 2. 8.

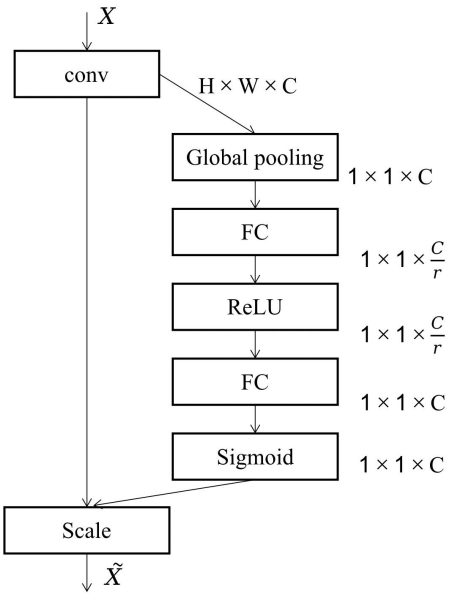

Fig. 2. 9. Diagram of SE block.

The diagram above shows a concrete implementation of the SE block, where the branch on the right is designed to calculate the weights of the individual channels. First, the input is processed through a global pooling layer, then two fc layers, and a sigmoid layer to obtain the final weights. The resulting weights are then multiplied with the input feature maps. After the processing, the neural network can adaptively learn the relative importance of different feature maps and thus focus more on the important ones.

2.5 Facial prior information

Unlike natural images, the subject of a face image is the human face, which has obvious structural features, the location and general shape of each person's facial organs are similar, and the face itself is endowed with some unique structural prior knowledge. Prior knowledge can provide structural information to the network, thus assisting in face image super-resolution. Broadly speaking, the structural prior knowledge of a face image typically consists of face heatmaps and face parsing maps,

etc.

There exists some face super-resolution methods that use prior information. FSRNet [16] uses face landmarks and face parsing maps to boost the performance of face image super-resolution. PFSR [17] uses a special training scheme to progressively generate high quality face images, and proposes a facial attention loss that gives higher weights to facial features near the facial landmark. EIPNET [18] generates high-quality face images by using a lightweight edge block to obtain edge information and an identity loss to constrain human identity information.

Chapter 3 Proposed methods

3.1 Proposed method 1

3.1.1 Model structure

The whole structure of the model is presented in Fig. 3.1. There are three main parts: the prior prediction branch, the super-resolution (SR) branch, and the fusion module. The structure and function of each part will be described in detail later. The input is a low resolution (LR) face image, the result of the SR branch is a coarse super-resolution (SR) image, the result of the prior prediction branch is predicted face parsing maps, and the output of the fusing module is the final super-resolution (SR) image. Each module has a corresponding loss function to optimize its parameters.

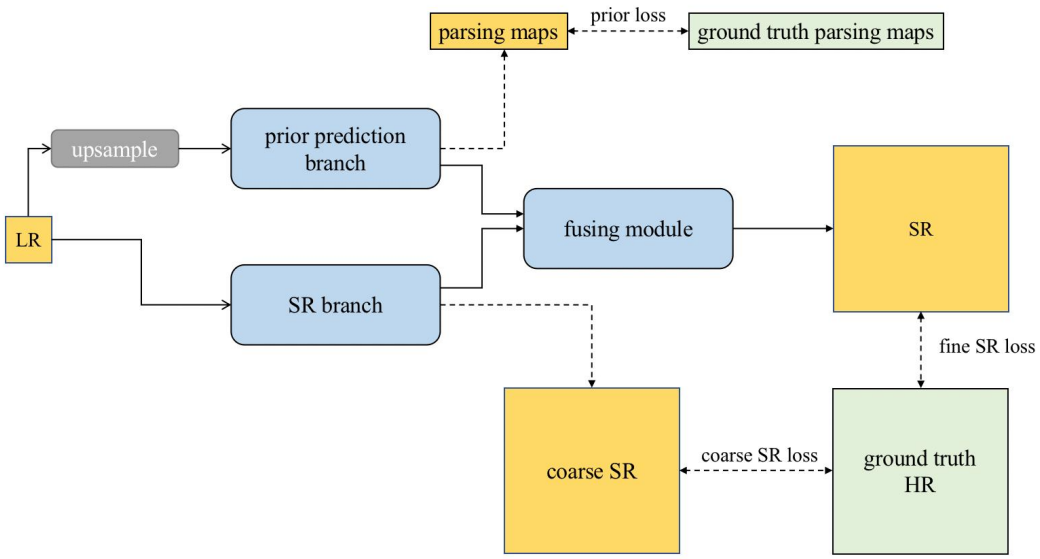

Fig. 3. 1. The network architecture of proposed method 1.

3.1.1.1 SR branch

In this branch, we adopt the structure of SRResNet and modify its res-block by introducing the channel attention mechanism. The whole framework is shown in Fig. 3. 2 and a more detailed structure is shown in Table 3. 1. Our basic block of SR

branch is shown in Fig. 3. 3 and the SR branch has 16 such basic blocks in total. We use convolution kernels of different sizes in the feature extraction part and fuse the features to strengthen the feature extraction capacity of the model.

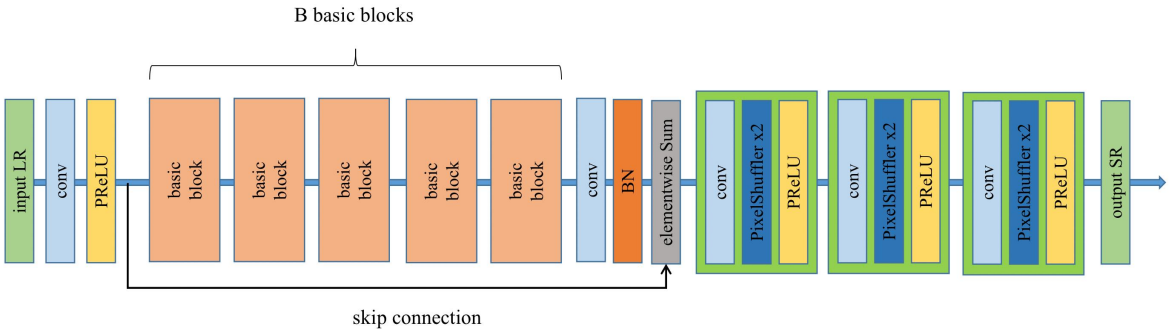

Fig. 3. 2. Structure of our SR branch.

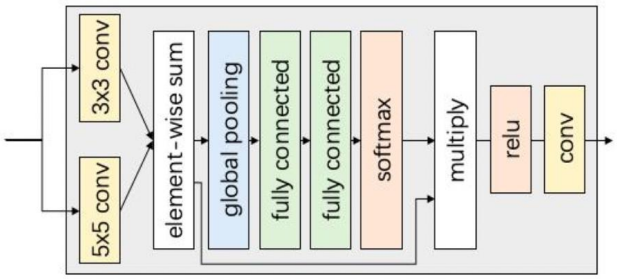

Fig. 3. 3. Structure of the basic block.

Fin is the input feature maps, $F_{3\times3}$ and $F_{5\times5}$ are the results of convolution operations with convolution kernels of 3 and 5 respectively, and F' is the fused result of $F_{3\times3}$ and $F_{5\times5}$. The process can be illustrated by equation 3.1:

$$F' = F_{3x3} + F_{5x5} \tag{3.1}$$

To achieve different sizes of convolution, we use the dilated convolution, which is shown in Fig. 3. 4. By introducing the dilation rate, the dilated convolution can bring a larger field of perception for the same size of convolution kernel. Accordingly, the dilated convolution has fewer parameters than the normal convolution with the same field size. For a standard convolution of size 3×3, the convolution kernel

contains a total of 9 parameters. During the convolution process, the elements in the convolution kernel are multiplied and summed pixel by pixel with the elements at the corresponding positions on the input matrix. Compared with the standard convolution, the dilated convolution has an additional parameter of dilation rate, which controls the distance between adjacent elements in the convolution kernel, and this parameter can control the size of the perceptual field. In Fig. 3. 4, the dilated convolutions with the same kernel size and different dilation rates are shown, from left to right, the dilation rate is 1, 2, 4, respectively.

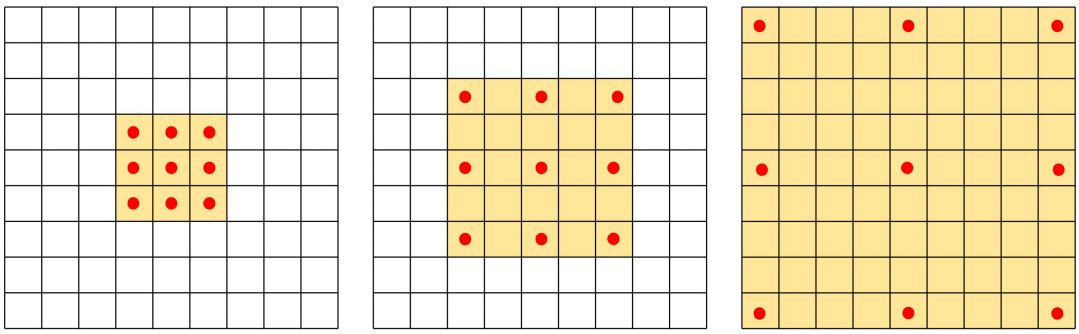

Fig. 3. 4. Dialated convolution.

A global pooling operation is then performed on F' and each channel is compressed into a single value, this process is called squeeze. Each value can represent its channel features. After this step, the feature map is compressed to a sequence of real numbers of size 1×1×C, which can be represented by equation 3.2.

$$z_c = GP(F') = \frac{1}{H*W}\sum_{i=1}^{H}\sum_{j=1}^{W} u_c(i,j) \tag{3.2}$$

where W and H are the horizontal and vertical size of U, correspondingly. U_c is the feature information of the cth channel, and Z_c is the value obtained after pooling for the c_{th} channel.

The next part is called the excitation and its main components are two dense layers and a softmax layer. As the squeeze operation acts on one layer of the feature map, the two fully connected layers are adopted to rearrange the feature information

between the channels and map the output values to between 0 and 1 using the softmax function. The output values are a series of real numbers of 1×1×C, each value representing the weight assigned to that channel, the target of SE block is to emphasize the useful channel features and ignore the useless ones, and the process can be represented by equation 3.3:

$$F_{scale} = s(fc_2(fc_1(z))) \tag{3.3}$$

where s represents the softmax activation function, fc_1 and fc_2 represent the first and second dense layers respectively, and F_{scale} is the output after softmax operation. The obtained F_{scale} is multiplied with F' and then passed through a ReLU function and a convolution layer to get the final output. This is shown in the following equation:

$$F_{out} = conv(\delta(F' * F_{scale})) \tag{3.4}$$

where F_{out} is the result of the SE block, conv is convolution operation and δ is ReLU function. Following SRResNet, we used the PixelShufffle layer to scale the feature maps to the target resolution.

Table 3. 1. Detailed network structure of SR branch.

operation	input dimension	output dimension
conv1	16×16×3	16×16×64
basic block 1- basic block 16	16×16×64	16×16×64
conv2	16×16×64	16×16×64
pixelshuffle 1	16×16×64	32×32×64
pixelshuffle 2	32×32×64	64×64×64
pixelshuffle 3	64×64×64	128×128×64
conv2	128×128×64	128×128×11

3.1.1.2 Prior prediction branch

Our prior prediction branch follows FSRNet [16], which contains several residual blocks and hourglass blocks. The input of the face prior branch is a low resolution image that has been up-sampled by Bicubic interpolation, the hourglass structure is adopted to obtain the parsing map of the face. LR face images provide limited information, it will be helpful to reconstruct the facial components if the accurate parsing map can be predicted and used as the prior information.

Table 3. 2. Detailed network structure of prior prediction branch.

operation	input dimension	output dimension
conv1	128×128×3	128×128×64
residual block 1	128×128×64	128×128×128
residual block 2-3	128×128×128	128×128×128
hour-glass block 1	128×128×128	128×128×128
hour-glass block 2	128×128×128	128×128×128
conv2	128×128×128	128×128×11

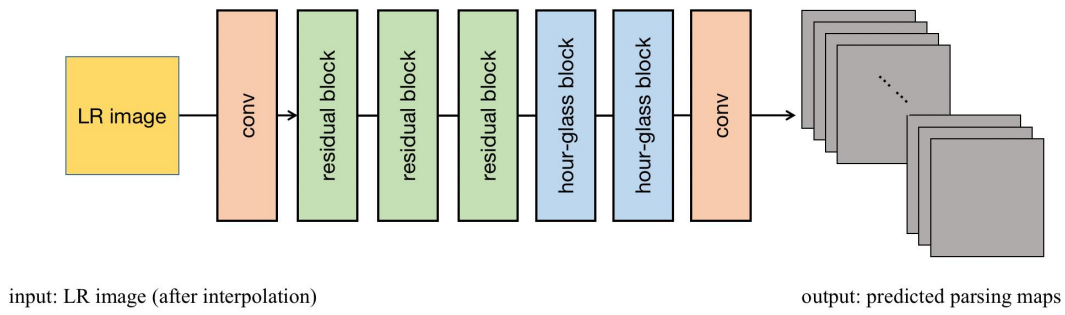

Fig. 3. 5. Prior prediction branch.

The structure of the prior prediction branch is shown in Fig. 3. 5, which contains three residual blocks and two hourglass blocks. The structure of the residual block is shown in Fig. 3. 6 and the purpose of it is feature extraction.

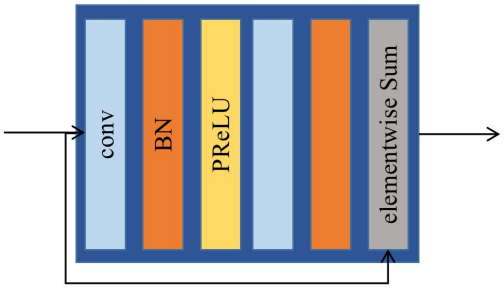

Fig. 3. 6. Structure of residual block.

Next we will introduce the hourglass block. Hourglass block is more frequently used in human pose estimation [19], the structure is shown in Fig. 3. 7. The feature map size will progressively decrease from the beginning to the middle part; from the middle to the end, the feature map size will gradually enlarge, this shape also forms the so-called hourglass structure. This structure can capture multi-scale features and provide richer information.

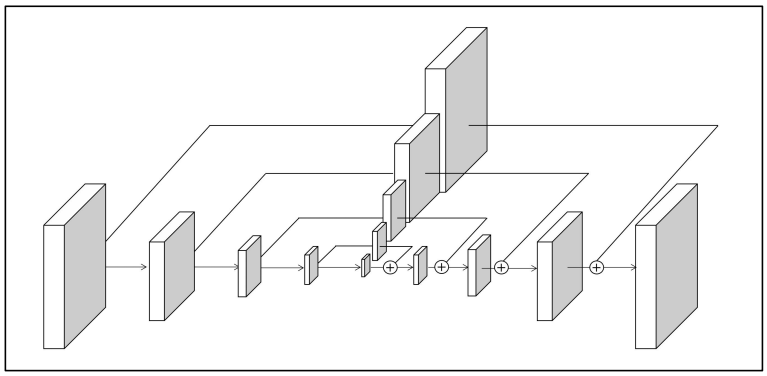

Fig 3. 7. The structure of an hourglass block.

3.1.1.3 Fusing module

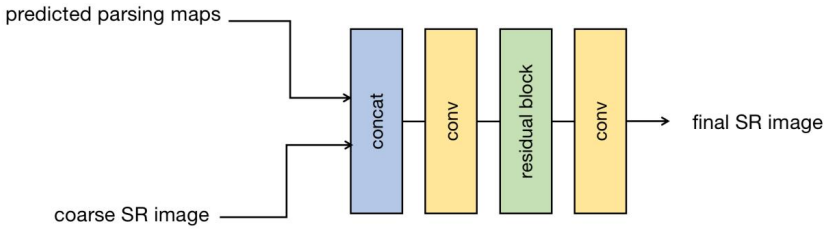

Fig. 3. 8. The fusing module.

The fusion module shown in Fig. 3. 8 is used to integrate the information from both branches for the final information integration and reconstruction, in which several conv layers and a residual block are utilized.

Regarding the fusion of two inputs, there are various methods that can be employed, here we use concatenation to fuse the features, which is done by stacking the inputs in the channel dimension. The stacked features are further integrated by subsequent convolutional layers. As shown in Table 3. 3, the detailed structure of the fusing module are presented.

Table 3. 3. Detailed network structure of fusing module.

operation	input dimension	output dimension
concat	128×128×3 and 128×128×11	128×128×14
conv1	128×128×14	128×128×64
residual block	128×128×64	128×128×64
conv2	128×128×64	128×128×3

It is worth mentioning that the super-resolution branch and the prior prediction branch are trained simultaneously.

3.1.2 Loss function

The total loss of this model contains three components: coarse SR loss, a prior loss, and fine SR loss.

coarse SR loss: MSE loss between coarse SR image and real HR image, which can be defined as:

$$L_c = \frac{1}{r^2WH}\sum_{x=1}^{rW}\sum_{y=1}^{rH}\left(I_{x,y}^{HR} - I_{x,y}^{coarse\_SR}\right)^2 \tag{3.5}$$

where I^{HR} is the HR face image, I^{coarse_SR} is the output of SR branch, r is the scaling factor, and W×H is the size of LR face image.

prior loss: MSE loss between predicted parsing maps and true parsing maps, which can be defined as:

$$L_p = \frac{1}{r^2WH}\sum_{x=1}^{rW}\sum_{y=1}^{rH}(P_{x,y}^{GT} - P_{x,y}^{predicted})^2 \quad (3.6)$$

where P^{GT} is the ground truth face parsing map, $P^{predicted}$ is the output of the prior branch, r is the scaling factor, and W×H is the size of LR face image.

fine SR loss: MSE loss between final SR image and ground truth HR image, which can be defined as:

$$L_f = \frac{1}{r^2WH}\sum_{x=1}^{rW}\sum_{y=1}^{rH}(I_{x,y}^{HR} - I_{x,y}^{final\_SR})^2 \quad (3.7)$$

where I^{HR} is the HR face image, I^{final_SR} is the output of the whole model, r is the scaling factor, and W×H is the size of LR face image.

The total loss called content loss $L_{content}$ is a sum of the above three losses.

$$L_{content} = L_c + L_p + L_f \quad (3.8)$$

3.2 Proposed method 2

In this method, we add GAN, which aims to generate more details of the face.

3.2.1 Model structure

It is well known that GAN contains two important parts: the generator and the discriminator. In this approach, the generator directly adopts the model proposed in method 1. And as for the discriminator, we use the discriminator in SRGAN, which contains multiple convolutional layers, the structure is shown in Table 3. 4. And the whole structure of the proposed method 2 is shown in Fig. 3. 9.

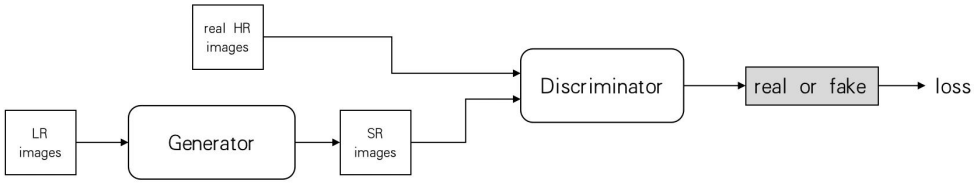

Fig. 3. 9. GAN model for super-resolution task.

Table 3. 4. Detailed network structure of discriminator.

operation	input dimension	output dimension
conv1	128×128×3	128×128×64
LReLU	128×128×64	128×128×64
conv block 1	128×128×64	64×64×64
conv block 2	64×64×64	64×64×128
conv block 3	64×64×128	32×32×128
conv block 4	32×32×128	32×32×256
conv block 5	32×32×256	16×16×256
conv block 6	16×16×256	16×16×512
conv block 7	16×16×512	8×8×512
Dense 1	8×8×512	1×1×1024
LReLU	1×1×1024	1×1×1024
Dense 2	1×1×1024	1×1×1
Sigmoid	1×1×1	1×1×1

3.2.2 Loss function

In related works we introduced that the training of GAN is a "min-max" process, and for our face super-resolution task this problem converts to

$$\min_{\theta_G}\max_{\theta_D} E_{I^{HR}\sim p_{train}(I^{HR})}[\log D_{\theta_D}(I^{HR})] + E_{I^{LR}\sim p_G(I^{LR})}[\log(1 - D_{\theta_D}(G_{\theta_G}(I^{LR})))]. \quad (3.9)$$

Since generators and discriminators are trained alternately in GAN, we elaborate the loss functions of generators and discriminators separately.

Generator loss:

$$L_g = \alpha L_{content} + \beta L_{vgg} + \gamma L_{adv\_g} \tag{3.10}$$

where $L_{content}$ is the loss introduced in proposed method 1, L_{vgg} is vgg loss, L_{adv_g} is the adversarial loss for generator, and α, β, γ are the weight factors for each loss.

The vgg loss can be obtained by the following two steps.

1) get the features extracted from the SR image and the real HR image at a certain layer in VGG19 [20]
2) calculate the distance between the above two features

The vgg loss is defined in equation 3.11:

$$L_{vgg/i,j} = \frac{1}{W_{i,j} H_{i,j}} \sum_{x=1}^{W_{i,j}} \sum_{y=1}^{H_{i,j}} (a_{i,j}(I^{HR})_{x,y} - a_{i,j}(G_{\theta_G}(I^{LR}))_{x,y})^2 \tag{3.11}$$

where $a_{i,j}$ is the feature obtained after passing through the i_{th} convolutional layer in VGG19 and before the j_{th} pooling layer.

The use of vgg loss reduces the gap between image features, improves the perceived similarity between SR images and real images, and brings high frequency details to the images.

The loss function that GAN has a guiding role in the optimization of the network, which optimizes the generator by adversarial loss. The adversarial loss is shown in equation 3.12:

$$L_{adv\_g} = \sum_{n=1}^{N} -log(D_{\theta_D}(G_{\theta_G}(I^{LR}))) \tag{3.12}$$

Discriminator loss:

The loss of the discriminator is equation 2.1, which is not repeated here.

Chapter 4 Experiments and results

4.1 Dataset

We use 28,000 images from CelebAMask-HQ Dataset [21] to train the model and 1000 images are used for testing. This dataset contains face images and their corresponding 11 face parsing maps, including jawline, eyebrows, eyes, nose, mouth and ears. The size of the HR face image is 128×128, and we down-scale the HR image by the factor of 8 to obtain the LR face image, whose size is 16×16.

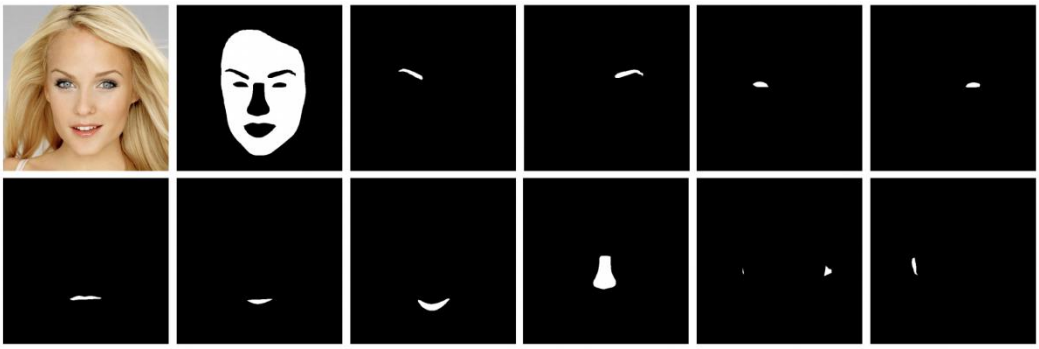

Fig. 4. 1. CelebAMask-HQ Dataset (image source: [21]).

4.2 Implementation details

For the optimizer, we use RMSProp, and we do the experiments using PyTorch framework on an NVIDIA GeForce RTX 1080 Ti.

4.3 Evaluation index

4.3.1 PSNR

In image/video compression and signal transmission, the PSNR is often utilized as a measure of the quality of image or signal and is defined below:

$$MSE = \frac{1}{pq}\sum_{i=0}^{p-1}\sum_{j=0}^{q-1}|X(i,j) - Y(i,j)|^2 \tag{4.1}$$

$$PSNR = 10log_{10}(\frac{MAX^2}{MSE}) \tag{4.2}$$

In measuring image quality, MSE is the mean square error between the two images, X(i, j) is a pixel value in image X, Y(i, j) is a pixel value in the reconstructed

image Y, and p and q indicate the number of pixels in the horizontal and vertical directions respectively. MAX represents the maximum possible grey scale value in the image and MAX equals to 255 when the image is quantized by 8 bits. Higher PSNR value usually means higher image quality with less distortion and noise.

4.3.2 SSIM

Structural similarity, often abbreviated as SSIM, is also an objective image quality evaluation metric that compares the similarity between images by considering brightness, contrast and structure, respectively.

$$l(x,y) = \frac{2\mu_x\mu_y + c_1}{\mu_x^2 + \mu_y^2 + c_1} \tag{4.3}$$

$$c(x,y) = \frac{2\sigma_x\sigma_y + c_2}{\sigma_x^2 + \sigma_y^2 + c_2} \tag{4.4}$$

$$s(x,y) = \frac{\sigma_{xy} + c_3}{\sigma_x\sigma_y + c_3} \tag{4.5}$$

L, c and s are the results of the comparison of luminance, contrast and structure respectively, $\mu_x = \frac{1}{N}\sum_{i=1}^{N} x(i)$ is the mean of x, $\mu_y = \frac{1}{N}\sum_{i=1}^{N} y(i)$ is the mean of y, $\sigma_x = \sqrt{(1/N-1)\sum_{i=1}^{N}(x(i)-\mu_x)^2}$ is the variance of x, $\sigma_y = \sqrt{(1/N-1)\sum_{i=1}^{N}(y(i)-\mu_y)^2}$ is the variance of y, $\sigma_{xy} = \frac{1}{N-1}\sum_{i=1}^{N}(x(i)-\mu_x)(y(i)-\mu_y)$ is the covariance of x and y, and C_1, C_2 and C_3 are constants. The formula for the total SSIM is shown in equation 4.6.

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{4.6}$$

The SSIM value is between 0 and 1, with larger values indicating that the structural distribution of the two images is more similar.

4.3.3 LPIPS

Learned Perceptual Image Patch Similarity (LPIPS), from [22], is a measure of the difference between two images, mainly the perceptual similarity between the SR image and the HR image. LPIPS is more consistent with human visual perception than traditional methods (e.g. PSNR, SSIM). LPIPS uses pre-trained deep networks (e.g. AlexNet [23], VggNet [20]) to extract deep image features and then measures the similarity of the images in the feature dimension.

4.4 Experiments and results

4.4.1 Objective evaluation

Table 4. 1. Objective evaluation result.

Method	PSNR(dB)	SSIM	LPIPS
BICUBIC	20.72	0.5200	0.5351
SRResNet	24.47	0.7201	0.1854
Proposed method 1	**24.66**	**0.7230**	0.1900
Proposed method 2	23.38	0.6689	**0.1018**

The objective result is shown in Table 4.1 , we can notice that the PSNR and SSIM of Bicubic are the lowest, and our proposed method 1 outperforms SRResNet in both PSNR and SSIM, with an increase of 0.19 db and 0.0029 in the two metrics, respectively. Compared with SRResNet and proposed method 1, proposed method 2 has lower PSNR and SSIM, but on the other hand has the lowest LPIPS value among all methods, which means better visual perceptual quality, and the performance will be presented by visualization in the next section.

4.4.2 Subjective evaluation

To show the performance of the proposed methods, we compare the qualitative results with SRResNet and the interpolation-based Bicubic method on the CelebmaskA-HQ face database.

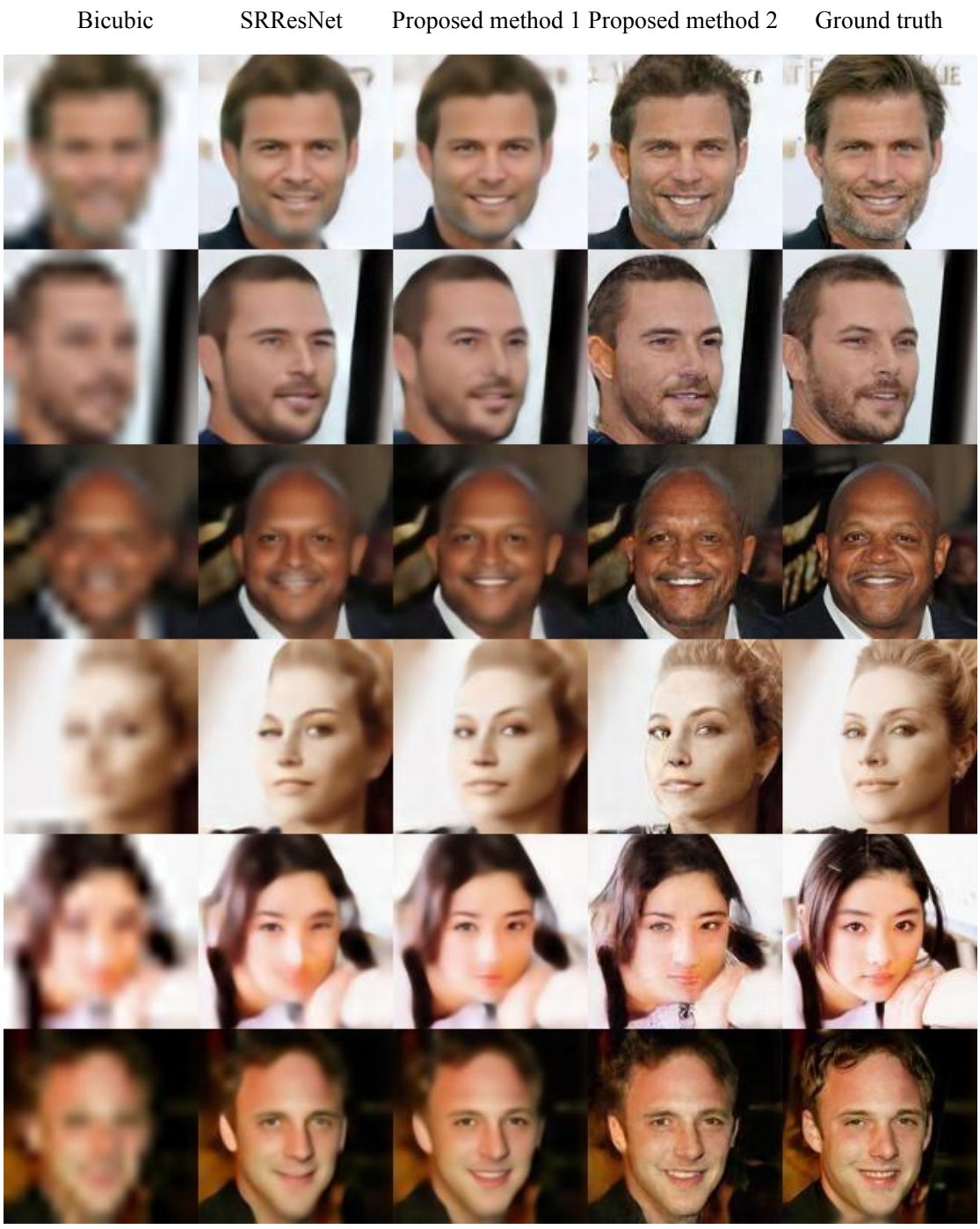

Fig. 4. 2. Qualitative evaluation result (image source: [21]).

As shown in Fig. 4. 2, the leftmost column shows the Bicubic results, followed by the results of SRResNet, proposed method 1 and proposed method 2, and Ground truth HR images. These images contain faces with different skin tones. Among them, the results obtained by Bicubic give a blurred appearance and have low visibility for the facial organs. As shown in the images of the fourth and fifth rows in Fig. 4. 2, the

human eye area obtained by the SRResNet method is blurred, and even the eyes seem to disappear. In addition, it can be seen from the images in the first, third, and sixth rows that SRResNet is weak in reconstructing the tooth parts and cannot accurately reconstruct the shape of the teeth. The proposed method 1 recovers the facial contours as well as the mouth and eye edges better. The proposed method 2 also has some improvement for the recovery of facial details. From the test images, we can see that both proposed methods have improved the super-resolution performance of the face.

Fig. 4. 3 compares the results of the method without GAN and with GAN, and we can notice that the method with GAN is able to generate more details of the face.

without GAN with GAN Ground Truth

Fig. 4. 3. Qualitative results with/without GAN (image source: [21]).

As can be observed from the first row, the GAN-based method generates a clearer eye area, and the outline of the glasses is more obvious; for the second row, the GAN-based method generates a face with clearer texture in the facial wrinkles and teeth part; similarly, the face in the middle of the third row has more realistic details in the eyebrows and beard part.

In addition to the above results, we will present some failure cases, as shown in Fig. 4. 4. One reason is that people may have accessories on their faces, such as glasses, and our model is still not very good at generating glasses. Another reason for failure is that when the expression of the mouth is not too conventional (e.g., laughing or duck face, etc.), the generated results will not be very accurate.

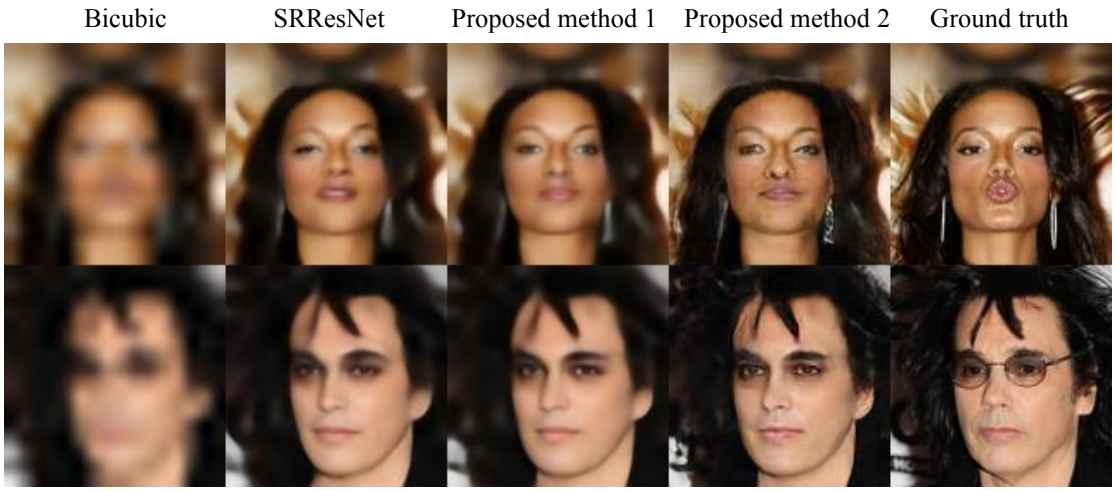

Fig. 4. 4. Failure cases (image source: [21]).

Chapter 5 Conclusion

We first introduce the research background of face image super-resolution, followed by the theoretical foundation of related techniques, including the basic theory of super-resolution, common super-resolution methods, and the basics of deep learning techniques used in this thesis. We propose two methods. The proposed method 1 is a CNN-based model, we add a multi-scale channel attention mechanism and face prior information based on SRResNet. The attention mechanism enables the deep model to be more focused on more useful channel feature information and ignore features that are less significant for the current task. The face prior can provide information on the location and contour of each organ of the face to assist in super-resolution. Our model contains three main parts, namely SR branch, prior prediction branch and fusing module.The three parts are optimized by their respective loss functions. The proposed method 2 is an extension of proposed method 1 and incorporates the training idea of GAN. Experiments reveal that the proposed method 1 has improved PSNR and SSIM values compared with SRResNet and Bicubic; although the PSNR and SSIM of our second proposal have decreased, its LPIPS is optimal, which means better subjective perception results for human eyes.

Chapter 6 Appendix

6.1 List of academic achievements

Yun Liu, Remina Yano, Hiroshi Watanabe, Takuya Suzuki, Takeshi Chujoh, and Tomohiro Ikai: “Text Image Super Resolution Using Deep Attention Neural Network,” IEEE Global Conference on Consumer Electronics (GCCE), OS-TMR(3), pp.306-308, Oct. 2021.

Yun Liu, Remina Yano, Hiroshi Watanabe, Takuya Suzuki, Zheming Fan, Takeshi Chujoh, and Tomohiro Ikai: “A Prior-Guided Face Image Super-Resolution Network Based on Attention Mechanism,” International Workshop on Advanced Image Technology 2023 (IWAIT 2023), No.7, Jan. 2023.

Bibliography

[1] H. Stark, P. Oskoui, “HR image recovery from image-plane arrays, using convex projections,” Journal of the Optical Society of America A-Optics & Image Science, 6(11): pp. 1715-1726,1989.

[2] M. Irani, S. Peleg, “Super resolution from image sequences,” 10th International Conference on Pattern Recognition, vol.2, pp. 115-120, 1990.

[3] J. Yang, J. Wright, T. Huang, Y. Ma, “Image super-resolution as sparse representation of raw image patches,” in 2008 IEEE conference on computer vision and pattern recognition, pp. 1-8, 2008.

[4] C. Dong, C. C. Loy, K. He, X. Tang, “Image super-resolution using deep convolutional networks,” in IEEE transactions on pattern analysis and machine intelligence, pp. 295-307, Oct. 2015.

[5] C. Dong, C. C. Loy, X. Tang, “Accelerating the super-resolution convolutional neural network,” in ECCV, pp. 391-407, Oct. 2016.

[6] J. Kim, J. K. Lee, K. M. Lee, “Accurate image super-resolution using very deep convolutional network.,” in proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1646-1654, Jun. 2016.

[7] W. Shi, J. Caballero, F. Huszar, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, Z. Wang, "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.

[8] T. Dai, J. Cai, Y. Zhang, S. T. Xia, L. Zhang, “Second-order attention network for single image super-resolution,” in IEEE/CVF CVPR, pp. 11065-11074, Jun. 2019.

[9] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, Y. Fu, “Image super-resolution using very deep residual channel attention networks,” in ECCV, pp. 286-301, Sep. 2018.

[10] Y. Zhang, Y. Tian,Y. Kong, B. Zhong, Y. Fu, “Residual dense network for image

super-resolution," in CVPR, pp. 2472-2481, Jun. 2018.

[11] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, "Photo-realistic single image super-resolution using a generative adversarial network," in proceedings of the IEEE conference on computer vision and pattern recognition, pp. 105-114, Jul. 2017.

[12] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley,S. Ozair, Y. Bengio, "Generative adversarial networks," Communications of the ACM, 63(11), pp. 139-144.

[13] K. He, X. Zhang, S. Ren, J. Sun, "Deep residual learning for image recognition," in proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778, 2016.

[14] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, C. C. Loy, "Esrgan: Enhanced super-resolution generative adversarial networks," in proceedings of the European conference on computer vision (ECCV) workshops, pp. 0-0, 2018.

[15] J. Hu, L. Shen, G. Sun, "Squeeze-and-excitation networks," in CVPR, pp. 7132-7141, Jun. 2018.

[16] Y. Chen, Y. Tai, X. Liu, C. Shen, J. Yang, "Fsrnet: End-to-end learning face super-resolution with facial priors," in proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2492-2501, Jun. 2018.

[17] D. Kim, M. Kim, G.Kwon, D. S. Kim, "Progressive face super-resolution via attention to facial landmark," arXiv preprint arXiv:1908.08239, 2019.

[18] J. Kim, G. Li, I. Yun, C. Jung, J.Kim, "Edge and identity preserving network for face super-resolution," in Neurocomputing, 2021.

[19] A. Newell, K. Yang, J. Deng, "Stacked hourglass networks for human pose estimation," in European conference on computer vision, pp. 483-499, Oct. 2016.

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

[21] http://mmlab.ie.cuhk.edu.hk/projects/CelebA/CelebAMask_HQ.html). Under the agreement that The CelebAMask-HQ dataset is available for non-commercial research purposes only.

[22] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, O. Wang, “The unreasonable effectiveness of deep features as a perceptual metric,” in proceedings of the IEEE conference on computer vision and pattern recognition, pp. 586-595, 2018.

[23] A. Krizhevsky, I. Sutskever, G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” Communications of the ACM, 60(6), pp. 84-90, 2017.