Film and Television Production AI Intelligent Optimization Processing System Based on Neural Optimization Algorithm

Dongsheng Yang*

{*Corresponding author: 17861013177@163.com}

School of Communication, Qufu Normal University, Rizhao 276800, Shandong, China

Abstract. In the context of the development of ultra high-definition (4K, 8K) industry, artificial intelligence technology represented by deep learning is developing rapidly in the field of image super resolution. Based on the adversarial generation super resolution network (SRGAN), we propose a novel image super resolution generation model combining semantic segmentation probability graph and iterative check kernel (IKC) technology. This model can recognize the target object in the image according to the application requirements and make the generated ultra HD image texture more real. Therefore, we made Images From TV (IFTV) data set based on the big data of radio and television media assets to optimize and train the common application scenes of radio and television (such as those with more faces or text), so that the model could achieve satisfactory super resolution effect in multiple scenes. It will provide strong support for the production of ultra HD content in the field of broadcasting and television in the future.

Keywords: Neural Optimization Algorithm, AI, Film and Television Production

With the development of ultra-high definition TV technology and the improvement of people's spiritual and cultural needs, the shooting technology and the video resolution presented by film and television works are becoming higher and higher. People not only get spiritual food from it, but also enjoy very good visual viewing experience. Traditional classic film and television works are still loved and concerned by the public today. However, due to the limitation of filming equipment and conditions at that time, as well as the damage caused by the long-term preservation of film and tape, the image quality of these works can no longer meet the needs of audiences in today's high-definition and ultra-high-definition era. In order to reproduce classic images, the film and television materials in the collection must be restored. And adopt new technology to improve video and audio quality.

For the materials with poor clarity, noise and jagged edges, how to make good use of these video resources to improve image clarity, reduce noise and remove jagged edges has become an urgent problem to be solved in front of video producers. Surprisingly, with the continuous development of AI image enhancement technology in recent years, some AI video processing software came into being, bringing new life

ICSETPSD 2023, November 17-18, Coimbatore, India

DOI 10.4108/eai.17-11-2023.2342814

to these old videos. Today, under the guidance of mobile Internet, big data, supercomputing and other new technologies, artificial intelligence technology is developing rapidly, especially in video processing, image recognition and enhancement, image understanding and analysis, and program content production and other aspects have achieved certain technical achievements. Advances in key technologies such as machine learning and computer vision processing have greatly promoted the development of AI technology and broadened the application field of AI in the media industry. At present, artificial intelligence technology combined with the development of 4K/8K ultra HD has been gradually applied to all links of the radio and television industry, assisting the production of programs, improving production efficiency and innovating content forms. The launch of 4K channels will certainly bring massive demand for ultra-high-definition programs. Except for a few new 4K original programs, most historical data and the standard definition and high-definition media assets of broadcasting stations need to match the high standard 4K broadcasting requirements through image quality improvement. The current mainstream image super resolution technology can support the improvement of video quality from standard definition to 4K (16x) or standard definition to HD (4x). Image super resolution technology can provide a "radiant" new face for these classic content materials.

1 Technical Background and Related Research

Image Super Resolution (hereinafter referred to as SR) is a classic computer vision topic, including single image Super Resolution, multiple image super resolution and video super resolution. Traditional SR technology is mainly based on frequency and spatial algorithms, including non-uniform interpolation method, iterative back projection method, maximum a posteriori probability method, etc., but the limitation is that it is only applicable to the case of spatially invariant noise and can only deal with the case that there is only global motion in the image but no local motion, and it is difficult to use prior information in the process of processing [1].

With the rise of the concept of artificial intelligence, deep learning technology has been widely used in the field of image super resolution. The image Super Resolution algorithm based on Convolutional Neural networks emerges. Li et al propose an effective and high-speed enhancement and restoration method based on the dark channel prior (DCP) for underwater images and video[2]. Huang et al. present an approach based on combining information from two different GANs, both of which generate a visual representation of unseen classes[3]. J S et al. proposed an ML algorithm that can analyze medical images and accurately detect them to reliably diagnose early cancer. The method mainly includes image enhancement, enhancement of medical images, then use discrete orthogonal Stockwell transform (DOST) for feature extraction, and support vector machine (SVM) classifier for classification[4].

2 System Design and Implementation

Based on the EDVR video superfraction algorithm, the system adjusts the network structure and optimizes the parameters according to the actual application requirements. After several iterations, a video superresolution generation model with excellent performance is constructed, and satisfactory image enhancement effect is achieved. In the training process, a set of data set with radio and television video image characteristics was made by ourselves, and after targeted learning, the automatic repair function of common video damage was realized.

2.1 Algorithm

(1) Semantic segmentation and spatial feature transformation

Semantic segmentation is originally a concept of image classification, but it can improve the image quality of SR image in some details in the field of super resolution combined with semantic segmentation technology. based on semantic segmentation can generate more natural pixels for different regions (for different categories) in SR images by classifying image pixels on the basis of SRGAN. According to this idea, we proposed to use ADE20K data set to train multiple categories of semantic segmentation system, and combined with the characteristics of data in the field of radio and television to optimize subdivided objects such as people, objects and scenes, so as to generate more realistic textures in SR images.

For a single LR image x, the goal of super resolution is to generate $y^{\wedge}$ of an HR image as close as possible to the real image y. Generally based on the direct method is the use of feedforward neural networks (CNN) learning to theta as the parameters of the function $G(x|\theta)$, as shown in formula (1).

$$y^{\wedge} = G_{\theta}(x) \tag{1}$$

Where, x is the input LR image. The Loss function loss is minimized by optimizing parameter θ, so that the generated SR image is more similar to the real HR image, as shown in formula (2).

$$\theta^{\wedge} = \arg\min_{\theta} \sum_{i} Loss\left(y_i{}^{\wedge}, y_i\right) \tag{2}$$

In terms of semantic segmentation, it is assumed that P is a set of conditional probability graphs for semantic segmentation, p_k represents the probability of class k in semantics, and K is the total number of semantic classifications, as shown in formula (3).

$$\psi = P = \left(p_1, p_2, \ldots, p_k, \ldots, p_K\right) \tag{3}$$

Spatial feature transformation (SFT) is essentially to generate modulation parameters related to spatial features through affine transformation. STF learns the modulation parameter pair $(\gamma,\ \beta)$ by a mapping function M, and the value of M is determined by the conditional probability $M(\psi)$ of semantic segmentation, as shown in formula (4).

$$(\gamma, \beta) = M(\psi) \tag{4}$$

(2) GAN and IKC

GAN algorithm based on deep learning technology is the main framework of our research project [6], and its generation model mainly relies on adversarial learning, as shown in formula (5).

$$\min_{\theta} \max_{\eta} IE_{I^{HR} \sim P_{HR}} \log D_{\eta}\left(I^{HR}\right) + IE_{I^{LR} \sim P_{LR}} \log\left(1 - D_{\eta}\left(G_{\theta}\left(I^{LR}\right)\right)\right) \tag{5}$$

In essence, GAN uses maximum likelihood estimation (MLE) in statistical concept to conduct confrontation training and learn the data distribution of target image pixels. When the data distribution of SR image generated by G is close to that of the real HR image, the purpose of generating a high resolution image that is fake and authentic is reached. In general, HR images are used as data sources in training and are subsampled by Bicubic [5] to generate LR images. HR-LR image pairing training can generate images conforming to the data distribution of the training set. However, HR uses a known Gaussian distribution kernel to generate LR pictures through Bicubic, but LR pictures in the real world belong to an unknown data distribution (Gaussian distribution kernel is unknown). We cannot guarantee that the data distribution followed by the training set of this model is similar to the real picture. According to BlindSR study [7] conducted by Gu et al., Chinese University of Hong Kong, if the data distribution between the training set and the actual application image is different, the resulting HR image will be too sharp or too blurry, severely affecting application performance.

Through IKC, Gaussian kernel parameters can be modified several times according to the initial generation results to generate more accurate SR images. IKC consists of a kernel estimation network P and a kernel verification network C. As shown in Fig.ure 1, for the predictor P, we use four convolution layers and a global mean pool layer, and leakyRelu is selected as the activation function. The convolution layer gives the estimation of kernel H in space and forms the estimation graph. Then, the global mean pool layer gives a global estimate by averaging over space.

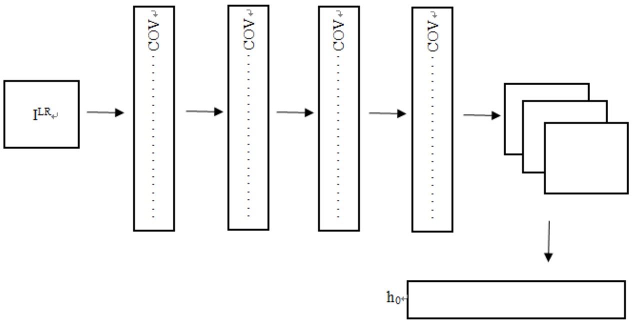

Fig. 1. Kernel estimation network

As shown in Fig.ure 2, the kernel verification network C consists of 5 groups of convolutional layers, and the estimated kernel h_{i-1} generated in the previous iteration is added as the input feature. h_{i-1} is extended by convolution to form F_h and stacked with F_{SR} to generate a new estimation graph, and Δh_i is given by global pooling operation.

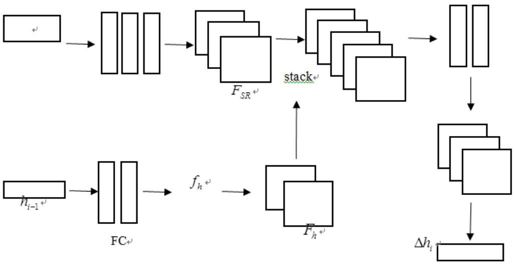

Fig. 2. Kernel verification network C

(3) Integration of semantic segmentation and IKC

Inspired by relevant studies [8], we combine GAN model based on semantic segmentation with IKC based on blind super resolution. According to the characteristics of its network structure, it can ensure the generation of SR images with more natural pixels for semantic segmentation of specific categories of images, and at the same time has the ability of self-checking fuzzy kernel, so as to maintain better performance for real world image super resolution.

We propose a new SFT module SFTconcat, which is derived from Fconcat by affine transformation of parameters (γ,β). As shown in FIG.. 3, Fsegcon is a classification probability graph generated by the semantic segmentation system and generated after specific convolution, which contains K-class object recognition categories. In order to avoid interference in different categories of an image, we try to reduce the convolution kernel size of semantic segmentation conditional network [15], and choose 1×1 convolution kernel to ensure that all kinds of pixels in the image are not affected.

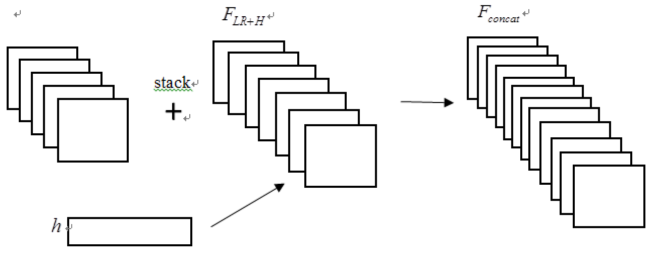

Fig. 3. The tensor construction of Fconcat

FLR is the feature vector graph generated by the input image through a specific convolution layer. Its structure is the three-dimensional tensor of C×H×W, where C is the number of channels, H is the height of the feature vector, and W is the width of the feature vector. H is developed from the estimated fuzzy kernel h, whose height and width are the same as H and W of FLR. The number of channels is b, which is obtained from fuzzy kernel h through PCA dimension reduction. Principal component analysis (PCA) [11] is derived from statistical algorithms. Its core function is to map feature data from high-dimensional space to low-dimensional space and preserve the relationship between feature values. Here, we assume that the size of the image fuzzy kernel is 1×1, and its vector space is 12. Through PCA, we map the 12-dimensional vector space to the B-dimensional linear space, take $M \in R^{b\times 1}$ as the mapping matrix, and through linear transformation h=Mk, $h \in R^{b}$, the value space of h is reduced from l2 to b, which greatly reduces the calculation cost while ensuring the validity [12]. Thus, the tensor size of FLR+H is (C+b)×H×W. The tensor shape of FLR and FLR+H has the same H and W dimensions, so the Fconcat stacked based on the C dimension can add new channel information while keeping the same spatial feature relation. Adding the predictive fuzzy kernel h to the SFT layer can make the network parameters affect the generation of ISR in the learning process, and thus play an important role in the subsequent iterative optimization.

As shown in Fig.ure 4, the spatial feature transformation diagram of SFTconcat is given. Fimg is the feature vector set (tensor) of the input image, and Fconcat is composed of Fimg, estimated fuzzy kernel H and semantic segmentation probability graph Fsegcon. Fconcat generated by specific convolution with Fimg tensor gamma and beta, with the same size by affine transformation $SFT_{concat}\left(F_{img}|\gamma,\beta\right)=\gamma\otimes F_{img}+\beta$. If the input image and segmentation probability graph are simply stacked and merged into the network input tensor, the performance of the entire network will have little effect. At the same time, GU et al. also pointed out [13] that when stacking tensors are directly input into deep CNN, the feature information of the first layer (such as fuzzy kernel information) cannot be transmitted to the deeper layer. After applying the spatial feature transform (SFT), the core map influences the output of the network by applying an affine transform to the feature map of each intermediate layer.

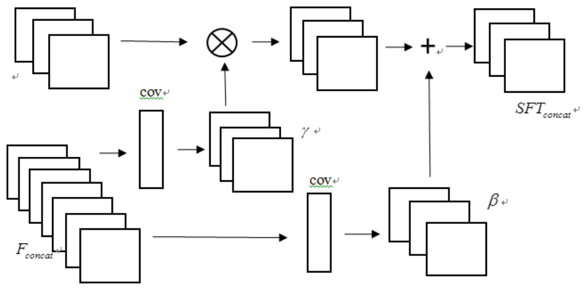

Fig. 4. Schematic diagram of the spatial feature transformation of SFTconcat

We use the residual module (RB) as the basic module to generate the model. As shown in Fig.ure 5, each RB consists of two SFT layers and two volume layers, namely, SFTconcat → COV → SFTconcat → COV. The application of RB makes the whole network have the excellent characteristics of Resnet, and effectively solves the problem of gradient disappearance in the training of deep neural network. The input image ILR passes through a series of convolutional layers and RB layers, and finally generates the ISR image of the target size after the feature is upsampled.

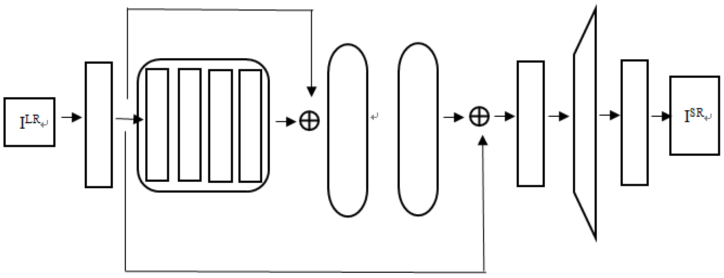

Fig. 5. Generation model G with SFTconcat as the basic unit

We take the 19-layer VGGnet with pre-training as the feature map corresponding to LR pictures to generate, and the corresponding Loss function is as follows.

$$Loss^{SR} = Loss_{vgg}^{SR} + 10^{-3} Loss_{GAN}^{SR} \qquad (6)$$

The discriminator network diagram is shown in Fig.ure 6. The discriminator network chooses Sigmoid function as the output layer, because the Sigmoid output range [-1,1] can better match the binary classification problem.

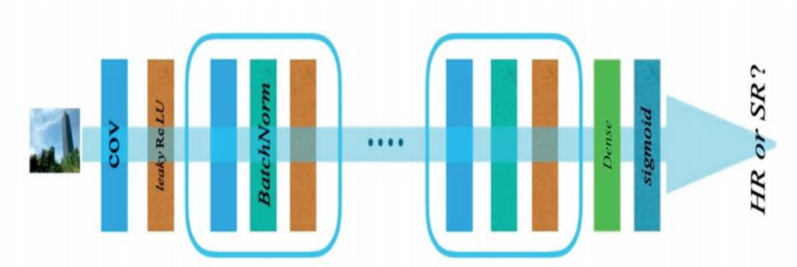

Fig. 6. Network diagram of discriminator D

2.2 Implementation

A deep learning project includes requirements analysis, model construction, training data preparation, parameter debugging, testing and other aspects. Only by combining each module can a whole project be realized. The super resolution system is developed for the production of program content in the field of radio and television. Due to the variety of content in TV programs, the specific pictures are also diverse. We only focus on optimizing the scenes and objects with high frequency, such as characters, characters, sky, animals, plants and water in the scenes of news, sports and entertainment programs. Therefore, for these goals, we will make targeted data sets and adjust the parameters. In terms of experimental environment, we equipped a piece of Nvidia Quadro M6000 and a piece of Nvidia Geforce 1080Ti to form a GPU array, and used Ubuntu16 operating system and Pytorch machine learning framework.

(1) Training data and preliminary preparation

Training data sets are the foundation of deep learning, the fuel of artificial intelligence. Therefore, selecting and crafting the right data set for the target is critical. It is usually necessary to divide the data set into three parts: the training set, the verification set (development) and the test set, with a ratio of 8:1:1. Most of the data is used for training, a little for development debugging and cross-validation, and the last bit for public testing. According to experience, the project in this paper conducts training and testing for 16 times image super resolution, namely WLR=WHR/4, HLR=HHR/4, ISR= ILR×16.

Firstly, COCO data set, which is widely used in the field of object detection, is selected as the pre-training data set of semantic segmentation model. COCO data set is a large general data set released by Microsoft for target detection, semantic segmentation and other training tasks. It contains more than 200,000 labeled pictures and various data of 80 categories of objects. The pre-trained semantic segmentation model based on COCO has the basic segmentation ability. According to the outdoor scene (OST) classification proposed by Wang et al. [14], eight common outdoor scenes (here we target at radio and television application scenes) are mainly classified by using the self-made IFTV data set to add the face and text classification.

The formula is shown below.

$$I^{LR} = \left(k \otimes I^{HR}\right)\downarrow_{DS} + n \qquad (7)$$

We believe that the HR image corresponding to all LR images can be found, and there is a specific fuzzy kernel k, so that the HR image can be restored to the corresponding LR image after downsampling through the kernel. The in-phase Gaussian fuzzy kernel was selected as the basis for generating the training data set, and the parameter of kernel width was selected as [0.2, 4] for the project with 16 times super resolution. For real natural pictures, there is often extra noise, so noise n is added, and n is assumed to obey Gaussian distribution and covariance σ =15. We adopted DIV2K and Flickr2K as HR data sets and obtained LR images by downsampling according to the above formula.

Similarly, DIV2K and Flickr2K were used as the basic data set, and IFTV data set was added for targeted tuning, and data was enhanced by means of flipping, symmetry, cropping, etc., so as to increase the data volume. According to the principle of machine learning, adding self-made radio and television media assets data can enhance the representation and generation ability of the model for specific scenes.

(2) Training and debugging

First, the semantic segmentation network is trained. Based on the trained segmentation model, ADE20K data set was used to optimize the classification of animals, plants and other outdoor scenes, and IFTV, a self-made data set, was used to optimize the classification of faces and characters that appear more frequently in TV programs. According to the statistical principle of machine learning, a similar data distribution between the training data and the actual data will make the performance of the training model better. Adding more scenes in the field of radio and television into the training set can effectively improve the classification accuracy of the model for similar scenes.

We used the IKC training data set made above to perform alternating training on SFT and IKC kernel prediction network and kernel verification network. The sub-training set (mini batch) contains the following items, where N indicates the size of mini-batch.

$$\{I_i^{HR}, I_i^{LR}, h_i\}, i \in [1, N] \tag{8}$$

According to IKC iteration formula, firstly train and update kernel prediction network P parameters, then update kernel verification network C parameters, and then update SFT module parameters in the subsequent iteration process. It is recommended that the number of iteration optimization is t=7, the size of mini-batch is 16, Adam optimizer is selected and super parameters β1=0.9, β2=0.999 are set, and the learning rate is lr=0.0001.

3 Test and Evaluation

By adding the generation model based on semantic segmentation probability graph to the ordinary SRGAN, the system can generate SR images of corresponding categories according to different regions in the image, and the material and details of the pixel are more consistent with the texture of the category. For example, the features of sky and water are obviously different. The semantic segmentation map will make the generated sky smoother and the water surface texture richer. As shown in Fig.ure 7, we can see that SR images generated by this system with super resolution are excellent in some details, especially in face and text. The face and text in the image are restored clearly, which is thanks to our use of IFTV data set. It is worth mentioning that the face details in SR image have also repaired the jagged texture caused by compression in the original image (Ground Truth). Through the SR model containing semantic segmentation, optimize the training of the objects such as face and text, the text and number in the SR Fig.ure are restored clearly, and even the jagged image contained in the original Fig.ure is repaired. Among them, the English and digital textures are clear and full, but there is still a certain gap between Chinese characters and the original image, which indicates that the number of Chinese character image samples in our self-made IFTV data set is not perfect enough and needs to be further improved.

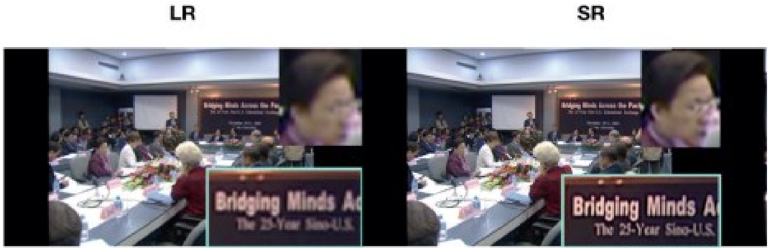

Fig. 7. Comparison results of test images

Usually, we will take PSNR and SSIM as the evaluation criteria of image processing quality. PSNR (peak signal-to-noise ratio) is a full-reference image objective standard, which takes the MSE of pixels between the reference image and the contrast image as the evaluation standard. SSIM, namely structural similarity, is also a full-reference image quality evaluation index, which measures the similarity between images with three dimensions of brightness, contrast and structure. In the application scenario where the superfraction of the model is 2X, 5 groups of test sequences are selected from the self-made feature data set to conduct full-reference objective quality assessment of the system. The Peak Signal to Noise Ratio (PSNR) and Structural Similarity (SSIM) corresponding to LR and SR images of each group before and after generation are shown in Table 1. It can be seen from the comparison that the video quality index has been improved after system processing.

Table 1. PSNR and SSIM Corresponding to SR and LR

	SET1 PSNR/SSIM	SET2 PSNR/SSIM	SET3 PSNR/SSIM	SET4 PSNR/SSIM	SET5 PSNR/SSIM
LR	23.8/0.83	26.3/0.88	18.3/0.68	17.2/0.68	18.50/0.70
SR	32.6/0.92	33.8/0.95	29.2/0.85	29.3/0.85	27.41/0.87

4 Conclusion

In this paper, we propose a method combining semantic segmentation graph and self-checking iterative kernel to generate SFT module through spatial feature transformation, and then join to form SRGAN network based on SFT module. In the process of development, starting from the practical application in the radio and television industry, we optimized the generation effect of specific scenes and images in the model training. For example, we added face and text, two categories widely existed in radio and television media assets, to the semantic segmentation map. To this end, we also use the big data of broadcasting and television's own media assets to establish IFTV data set, so as to provide data basis for the intelligent development of broadcasting and television media assets.

The prediction of LR image blur kernel in this paper was mainly aimed at Isotropic Kernels, while IKC could not completely solve the dynamic blur problem of images in practical application, which needed to be improved by subsequent research. On the other hand, the category of semantic segmentation in this project is limited to common scenes. If there are other categories in the image, super resolution will be carried out in the default way, and special optimization is not possible. In the future, we will explore more classification and recognition algorithms, and combine them with super resolution algorithms to subdivide and optimize more broadcast application scenarios. At the same time, the efficiency of the super resolution system is also a concern. As the project adopts the algorithm combining semantic segmentation, IKC and SRGAN, there are multiple neural networks in the system and they need to be trained separately or alternately. One-click learning has not been achieved yet. In the future, we plan to integrate this algorithm into a set of neural networks based on End2End mapping, and explore the unity of performance and efficiency in practical application.

References

[1] Chengda L,Xiang D,Yu W, et al. Enhancement and Optimization of Underwater Images and Videos Mapping.[J]. Sensors (Basel, Switzerland),2023,23(12).

[2] Kaiqiang H,Luis P M,Susan M. Enhancing Zero-Shot Action Recognition in Videos by Combining GANs with Text and Images[J]. SN Computer Science,2023,4(4).

[3] S J N,PALAKUZHIYIL V G. Colorectal polyp detection in colonoscopy videos using image enhancement and discrete orthonormal stockwell transform[J]. Sādhanā,2022,47(4).

[4] Hao S,Jun Y,Hongbo L. An image enhancement approach for coral reef fish detection in underwater videos[J]. Ecological Informatics,2022,72.

[5] Wang X, Chan K, Yu K, et al. EDVR: Video Restoration with Enhanced Deformable Convolutional Networks[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE, 2019.

[6] Zhou B, Hang Z, Fernandez F X P , et al. Scene parsing through ADE20K dataset[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017.

[7] Jianfeng Z,Kehong G,Xinchao W, et al. Learning to Augment Poses for 3D Human Pose Estimation in Images and Videos.[J]. IEEE transactions on pattern analysis and machine intelligence,2023,PP.

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

[9] He K, Zhang X, Ren S , et al. Deep Residual Learning for Image Recognition[C]// IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2016.

[10] Efrat N, Glasner D, Apartsin A , et al. Accurate Blur Models vs. Image Priors in

Single Image Super-resolution[C]// IEEE International Conference on Computer Vision. IEEE, 2013.

[11] Mathieu M, Couprie C, Lecun Y . Deep multi-scale video prediction beyond mean square error[C]// ICLR. 2016.

[12] Keys R G . Cubic convolution interpolation for digital image processing[J]. IEEE Transactions on Acoustics, Speech, and Signal Processing, 2003, 29.

[13] Gatys L A, Ecker A S, Bethge M , et al. Controlling Perceptual Factors in Neural Style Transfer[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017.

[14] Wang X, Yu K, Dong C , et al. Recovering Realistic Texture in Image Super-resolution by Deep Spatial Feature Transform[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2018.