

Article

Generative Adversarial Network-Based Distortion Reduction Adapted to Peak Signal-to-Noise Ratio Parameters in VVC

Weihaio Deng and Zhenglong Yang *

School of Urban Railway Transportation, Shanghai University of Engineering Science, Shanghai 201620, China; ycsd98@gmail.com

* Correspondence: 10190004@sues.edu.cn

Abstract: In order to address the issues of image quality degradation and distortion that arise in the context of video transmission coding and decoding, a method based on an enhanced version of CycleGAN is put forth. The lightweight attention module is integrated into the residual block of the generator module structure, thereby facilitating the extraction of image details and motion compensation. Furthermore, the perceptual function LPIPS loss is increased to align the image restoration effect more closely with human perception. Additionally, the network training method is modified, and the original image is divided into 128×128 small blocks for training, thus enhancing the network's accuracy in restoring details. The experimental results demonstrate that the algorithm attains an average PSNR value of 30.1147 on the publicly accessible YUV sequence dataset, YUV Trace Dataset, which is a 9.02% enhancement compared to the original network. Additionally, the LPIPS value reaches 0.2639, representing a 10.42% reduction, and effectively addresses the issue of image quality deterioration during transmission.

Keywords: video coding; image reconstruction; GANs; PSNR; LPIPS



Citation: Deng, W.; Yang, Z. Generative Adversarial Network-Based Distortion Reduction Adapted to Peak Signal-to-Noise Ratio Parameters in VVC. *Appl. Sci.* **2024**, *14*, 11561. <https://doi.org/10.3390/app142411561>

Academic Editor: Pedro Couto

Received: 22 October 2024

Revised: 3 December 2024

Accepted: 6 December 2024

Published: 11 December 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

H.266/VVC [1], a new video coding standard, offers higher compression rates than H.265/HEVC while maintaining quality. It supports 8K Ultra HD (shown in Figure 1), HDR content, and 360-degree video [2] and is optimized for streaming services and real-time communications that require dynamic bandwidth and resolution adjustments. VVC improves compression efficiency, compatibility, and flexibility for emerging video formats and application scenarios.

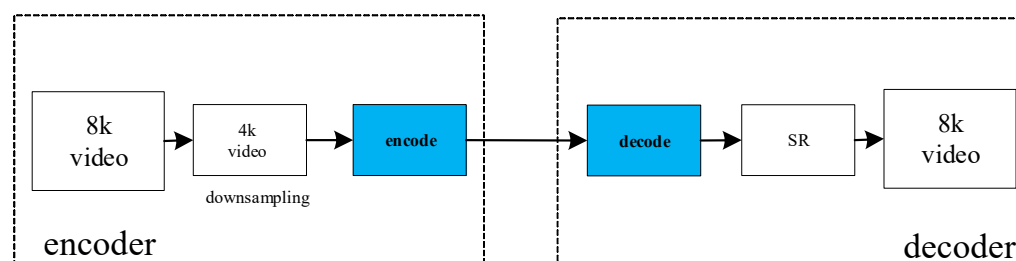


Figure 1. 8K video transmission system.

Video coding [3] improves storage and transmission efficiency by reducing redundant information. It uses prediction techniques to analyze similarities between frames (inter-frame) or within frames (intra-frame), leaving 'residuals'. These residuals are orthogonally transformed into the frequency domain, producing coefficients that represent different frequency components. Quantization simplifies these coefficients for storage and transmission, with the quantization parameter (QP) [4] determining the degree of simplification

(shown in Figure 2). A higher QP results in less accurate quantization, which reduces the amount of data and bandwidth required but compromises quality. Conversely, a lower QP increases data volume but improves video quality. Thus, video coding reduces data volume through prediction and quantization and balances video clarity and data requirements by adjusting the QP.

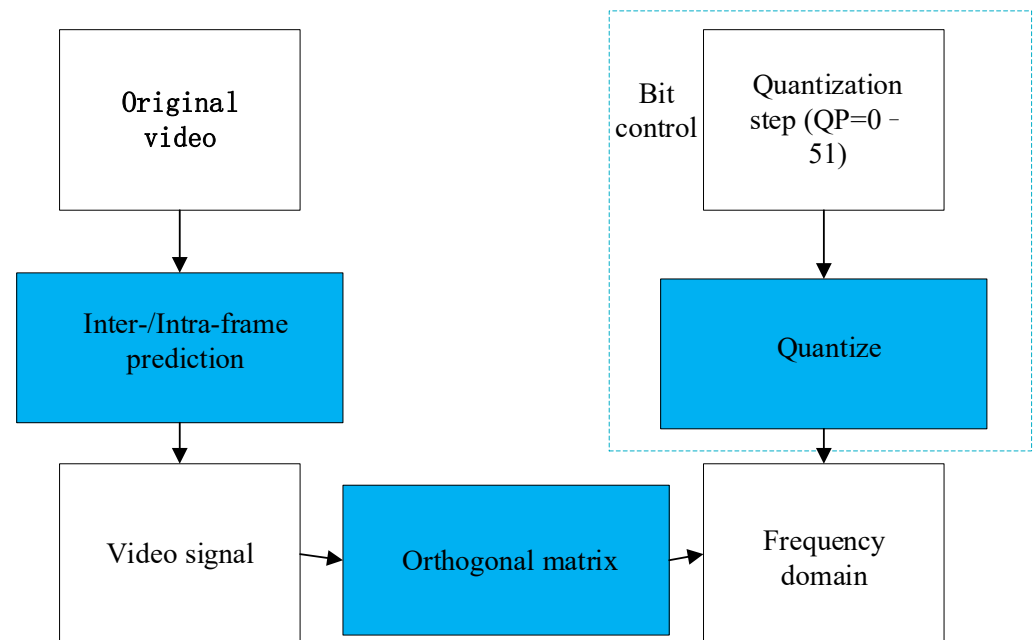


Figure 2. Control of video decoding by quantization parameters.

Video encoding and decoding are critical to video transmission, with the quantization parameter (QP) controlling data transmission and affecting picture quality. The video signal is de-redundant before coding and then converted to the frequency domain, where the coefficients are quantized and transmitted, resulting in a loss of quality. A number of potential solutions have been continuously proposed to improve this problem. The existing methods can be grouped into three dominant categories: interpolation-based, reconstruction-based, and learning-based.

The fundamental principle of interpolation is to estimate the value of an unknown pixel point by employing a numerical relationship between known pixel points. This process is based on the concept of signal processing, whereby the data from neighboring pixels is utilized to make inferences and thereby obtain a restored image. The most basic form of interpolation is linear interpolation. In this method, the pixel value is assumed to be a linear function of the position between two neighboring pixels, A and B. This allows the interpolated pixel value to be calculated using the pixel value and the position between A and B. Accordingly, the pixel value at any position between A and B can be calculated based on this linear relationship. Higher-order interpolation methods, such as bilinear and bicubic interpolation, consider additional information about neighboring pixels. Bilinear interpolation considers the linear relationship in both directions, building upon the principles of linear interpolation. In the case of a pixel situated within a rectangular region, the pixel values at any given position within the region can be calculated by linear interpolation in two directions based on the pixel values of the four vertices. Bicubic interpolation introduces a greater number of pixel points than bilinear interpolation, which takes into account information from a larger range of pixel values, thus improving the accuracy of the interpolation. The method typically employs more complex weighting functions.

A reconstruction-based algorithm is a type of image reconstruction that introduces a priori knowledge during the reconstruction process to achieve a locally optimal solution. This involves the use of frequency domain, time domain, and other multi-scale mapping

relationships to predict the results of the newly generated image. The most commonly used reconstruction-based algorithms are projection onto convex sets (POCS), maximum a posteriori (MAP) estimation, Bayesian analysis (BA), iterative back projection (IBP), and iterative back projection (IBP). Despite the addition of constraints to the image super-resolution reconstruction process through the introduction of a priori information, these methods yield satisfactory reconstruction results. However, these algorithms are afflicted by significant shortcomings, including suboptimal convergence. In certain application scenarios, they are even outperformed by the results of interpolation methods.

The current state-of-the-art algorithms are based on learning methods. Dong et al. [5] were the first to propose a convolutional neural network super-segmentation model, SR-CNN, which uses only a three-layer neural network (CNN) to perform super-resolution reconstruction. This method demonstrates superior performance compared to the traditional SISR method, proving that neural networks are well-suited for image restoration and enhancement. Subsequently, a rapid super-resolution network model (FSRCNN) was developed, employing up-sampling and inverse convolution to reconstruct high-resolution images. This approach reduced the computational complexity and enhanced the performance of the network. Kim et al. [6] proposed Very Deep Super-Resolution (VDSR), based on VGG-net, which increased the depth of the network to 20 layers and introduced skip-connectivity. The VDSR network, based on the VGG-net architecture, increases the depth of the network to 20 layers and introduces skip connections to enhance the stability of the training process and improve the quality of the reconstructed images. The enhanced deep super-resolution network (EDSR) eliminates the batch normalization (BN) layer from ResNet and increases the number of residual blocks and convolution kernels. Deep Back Projection Networks (DBPNs) incorporate the iterative back projection (IBP) algorithm into deep learning. Very deep residual channel attention networks (RCANs) employ a channel attention mechanism to optimize network channels, prioritizing those containing crucial details. The phenomenon of distortion following image encoding can be mitigated by super-resolution reconstruction using convolutional neural networks (CNNs) [7]. However, CNNs face two problems: quality degradation depends on image characteristics, which affects model performance, and CNNs struggle with transmission distortion. Image distortion in video codecs is typically caused by compression, transmission, and decoding errors and manifests itself as loss of detail, blurring, mosaicking, or color anomalies. The use of efficient codecs and optimized parameters can reduce the distortion, but deep learning methods based on Generative Adversarial Networks (GANs) can be used to repair the image distortion and improve the video quality prior to codec output.

GAN [8] is a deep learning framework that generates new data through adversarial competition between generators and discriminators. Its structure is shown in Figure 3. The generator mimics real data patterns and creates fake data with the goal of making it impossible for the discriminator to distinguish between the data it generates and the real data. The discriminator, on the other hand, recognizes the authenticity of the data and scores the real data high and the fake data low. During the training process, the generator and the discriminator continuously optimize their strategies and eventually reach a balance where the generator data are of extremely high quality and the discriminator cannot distinguish them.

However, a single GAN can lead to a homogenized style of image reduction, which may not reflect the characteristics of the whole dataset in a diverse way. The challenge of image reduction is, therefore, to produce images that faithfully reflect the details of the content data and also show the diverse stylistic features in the dataset rather than a single copy of a particular style. This article proposes an adaptive image enhancement method that adjusts processing strategies based on the degradation characteristics of the image coding process, thereby improving reconstruction quality.

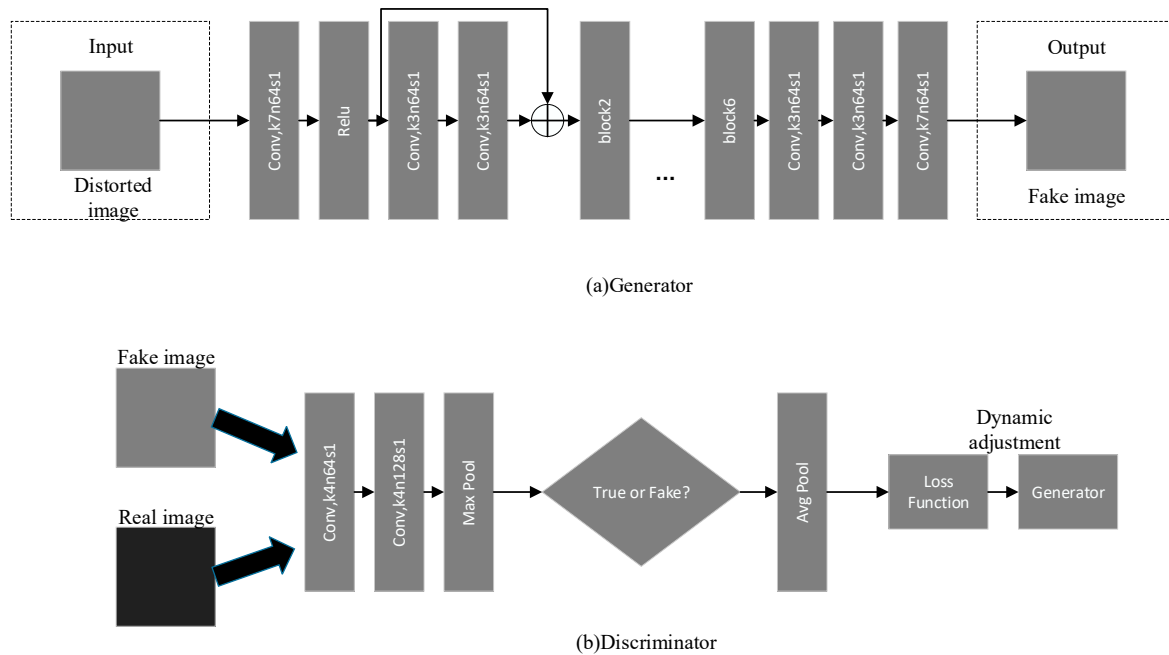


Figure 3. Basic structure of GAN network.

2. Proposed Algorithm

2.1. CycleGAN

The CycleGAN [9] model employs the adversarial training strategy of traditional GAN, innovating by incorporating a pairwise learning mechanism. This enables the model to achieve effective mapping between source and target domains in the absence of paired samples, thus eliminating the need for paired datasets for tasks such as style migration. This greatly expands the scope of application of the model.

In comparison with the conventional GAN model, the CycleGAN system employs a dual GAN network configuration, as illustrated in Figure 4. This architectural approach entails the transformation of the initial single generator and discriminator structure, thereby introducing a more intricate cyclic adversarial model to the system. The distorted image is initially processed by the GAN1 generator for restoration, after which it is presented to the discriminator for comparison with the real image. Following this, the trained realistic image is fed to the GAN2 generator once more to simulate the distortion and then subjected to another round of comparison with the distorted image.

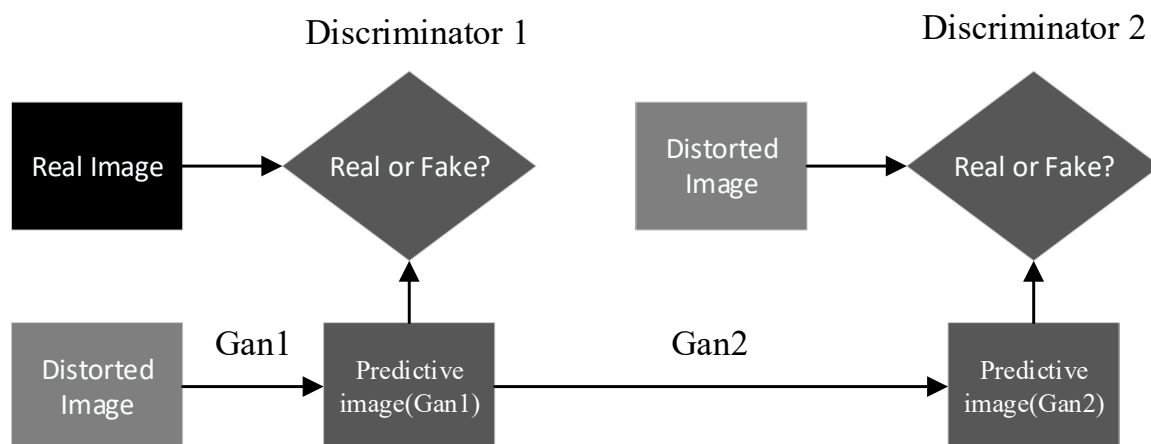


Figure 4. The CycleGAN network adversarial process.

The symmetric ring structure of CycleGAN is designed in such a way that it can be regarded as a system consisting of two Generative Adversarial Networks, which are in constant interaction with each other and with the distorted images that they are tasked with restoring. This structure not only enhances the flexibility of the model but also provides it with strong support in dealing with the task of image distortion repair without pairwise data.

2.2. The Generators Incorporate Lightweight Attention Mechanisms

The objective of optimizing video transmission is to facilitate the reception of a superior visual image. Although CycleGAN has demonstrated the capacity to reproduce image details with greater precision, this article proposes the integration of a lightweight attentional mechanism within the residual layer of the GAN generator as a means of enhancing the perceptual quality perceived by the human eye. The network structure is illustrated in Figure 5.

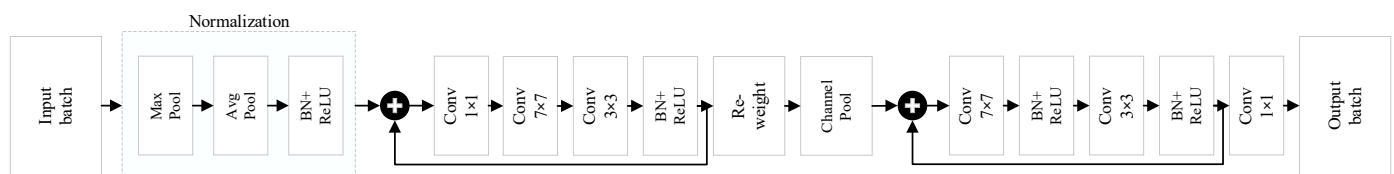


Figure 5. Lightweight attention mechanism.

The attention mechanism comprises two distinct components. The first part is to enhance the feature processing based on the spatial distribution of its own pixels after normalized feature extraction of the single-frame image; the second part is to compensate for the spatio-temporal action information by combining the images before and after the video frame in which the image is located, after channel pooling.

Firstly, single-frame images are batch normalized after maximum pooling as well as average pooling. This process transforms the input data at each level into a distribution with the same mean and variance, typically a standard normal distribution (mean 0, variance 1). This approach has been shown to accelerate training, enhance model convergence, and diminish the model's sensitivity to initialization and learning rate choices while simultaneously reducing internal covariate bias. Subsequently, the image is subjected to a ReLU (rectified linear unit) function, which is a prevalent activation function, particularly in convolutional neural networks within the domain of deep learning. In such contexts, the primary function of the activation function is to introduce non-linearities, thereby enabling the model to learn and simulate intricate function mapping relationships. In the absence of an activation function, irrespective of the number of layers in a neural network, the model is ultimately equivalent to a linear model, which would significantly restrict its expressive power. Gradient vanishing is a common issue where the gradient diminishes rapidly with the increase in layers during backpropagation, resulting in the weights in deeper layers of the network being updated very slowly. The derivative of the ReLU function is a constant 1 in positive intervals, which helps to alleviate the problem of vanishing gradients. The ReLU function is very simple to compute, with no complex mathematical operations, which makes it highly efficient in practical applications. The sparsity of the ReLU function also serves to mitigate the phenomenon of overfitting to a certain extent. Next, the image passes through a combination of 1×1 , 7×7 , and 3×3 , in turn containing convolutional kernels of different sizes. This is to capture features at different scales in the image. The 1×1 convolutional kernel is primarily responsible for capturing local details and texture information. Due to its small receptive field, the 1×1 convolutional kernel is capable of efficiently handling minor alterations in the image. The 3×3 convolutional kernel is primarily responsible for capturing medium-range features, such as edges and corners, thereby facilitating the extraction of distortions at the image edges with greater detail. The 7×7 convolutional kernel has the capacity to encompass a more expansive

receptive field, thereby facilitating the capture of a broader range of contextual information. This is advantageous for the comprehension of global structures and patterns within an image. The combination of convolutional kernels of varying sizes enhances the adaptability of the model, enabling it to accommodate disparate input dimensions and feature scales. This adaptability is crucial for processing images with diverse resolutions and scales. The end of the first part is completed by a residual structure, which serves to facilitate more profound network training while simultaneously addressing the issue of gradient vanishing and mitigating model degradation. This, in turn, serves to enhance the quality of the model representation.

The second part is initially processed through a Re-weight layer, which assigns varying degrees of importance to the extracted features, with a particular focus on the distorted portion of the image. This is performed to enhance the overall performance and facilitate the restoration of the distortion. Subsequently, a channel pooling operation is conducted, which represents a distinct type of pooling operation that differs from conventional spatial pooling techniques (e.g., maximum pooling and average pooling) in that channel pooling is executed in the channel dimension as opposed to the spatial dimension. The objective of this pooling operation is to reduce the number of channels in the feature map, thereby reducing the number of parameters and the computational burden in subsequent layers. This results in a more complex and computationally efficient model. The pooling of features from disparate channels enables the fusion of information from different channels, thereby facilitating the extraction of more representative features. The application of channel pooling serves to diminish the model's reliance on specific channel features, fortify its resilience to input alterations, and augment its capacity for generalization. Channel pooling captures the interchannel correlation, enabling the model to assign disparate weights across channels and thus enhance the representativeness of the features. Furthermore, the reduction in the dimensionality of the features that channel pooling entails also mitigates the risk of overfitting the model. The subsequent operations are in accordance with the preceding phase, thus ensuring uniformity in the training of the model.

Figure 6 illustrates the structure of the generator network following the addition of the attention mechanism.

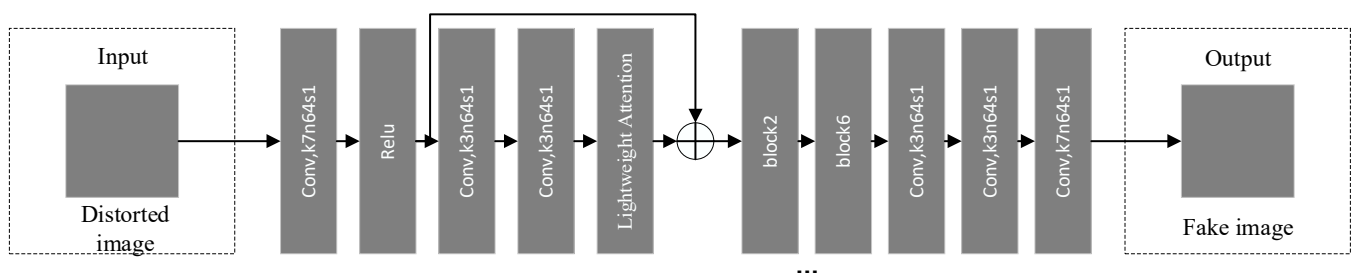


Figure 6. Lightweight attention mechanism.

2.3. Adjustment of the Discriminator Loss Function

The loss function of CycleGAN consists of the following three main components:

1. Generative Adversarial Network Loss is employed to train the generator to generate an image that can deceive the discriminator into believing that the generated image is authentic. This loss function is generic to the basic Generative Adversarial Network (GAN) and can be expressed for the generator G and the discriminator D_Y as follows:

$$L_{GAN}(G, D_Y, X, Y) = E_{y \sim p_{data}(y)} [\log D_Y(y)] + E_{x \sim p_{data}(x)} [\log(1 - D_Y(G(x)))] \quad (1)$$

where X represents the set of original images, Y represents the set of attempted generated images, $P_{data}(x)$ and $P_{data}(y)$ represent the probability distributions of the original and real images, and E and $\log$ are the mathematical expectation and logarithm, respectively.

2. Cyclic consistency loss, a loss function specific to CycleGAN, is employed to guarantee that the conversion of an image between two domains can be reversed. That is, if an image is converted from domain A to domain B and then converted from domain B back to domain A, it should be as close as possible to the original image. The cyclic consistency loss can be expressed as follows:

$$L_{cyc}(G, F) = E_{x \sim P_{data(x)}} [\| F(G(x)) - x \|_1] + E_{y \sim P_{data(y)}} [\| G(F(y)) - y \|_1]. \quad (2)$$

where F stands for the generator F that is opposite to the process of generator D, $\| \cdot \|_1$ is the number of paradigms, which is used to calculate the absolute value difference between the two images, and the other parameters are the same as in Equation (1).

3. Identity Loss: This loss function is employed to encourage the generator to preserve the original identity of the image on domains that do not necessitate transformation. In other words, when the input image is already part of the target domain, the generator should endeavor to output the same image as the input. The identity loss can be expressed as follows:

$$L_{id}(G, F) = E_{y \sim p_{data(y)}} [\| G(y) - y \|_1] + E_{x \sim P_{data(x)}} [\| F(x) - x \|_1]. \quad (3)$$

In order to target the training of the human eye on the perceptual effect of the reduced image, this article proposes the addition of the perceptual function LPIPS (Learned Perceptual Image Patch Similarity) after the traditional loss function. LPIPS [10] is a metric used to evaluate image similarity, which employs a deep learning model to assess perceptual differences. This metric is based on the assumption that even if two images are identical at the pixel level, a human observer may perceive them as different. Consequently, LPIPS employs pre-trained deep networks (e.g., VGG [11] and AlexNet [12]) to extract image features, subsequently computing the distance between these features to evaluate the perceptual similarity between images. This approach is more consistent with human perception than traditional pixel-level evaluation methods (e.g., PSNR [13], SSIM [14], and FSIM [15]), which are defined as follows:

$$L_{LPIPS}(X, Y) = \frac{1}{N} \sum_{i=1}^N \| \phi_i(X) - \phi_i(Y) \|_2. \quad (4)$$

The ultimate total network training loss function can be expressed as follows:

$$L_{CycleGAN} = L_{GAN} + L_{cyc} + L_{id} + L_{LPIPS}. \quad (5)$$

2.4. Network Training Program Improvement

A fundamental concept in the transmission process of video codecs is the quantization parameter (QP). The QP determines the fineness of the quantization. A higher QP value corresponds to a lower level of quantization precision; conversely, a lower QP value corresponds to a higher level of quantization precision. However, the amount of data is reduced in accordance with the QP value, which helps to reduce the bandwidth required for storage and transmission. Conversely, a reduction in the QP value will result in an increase in the precision of quantization, accompanied by an increase in the amount of data. However, this will lead to an improvement in the quality of the video. In essence, the QP allows for the calibration of video clarity and data consumption. Accordingly, in the context of training schemes for video coding, the dataset is typically partitioned based on the QP values observed prior to and following the coding process.

Even if the images are encoded with the same QP, the degradation will be different for each image. As a result, CycleGAN did not converge on every QP. For coded images, PSNR and image frequency distribution are related. For low-frequency images, the number of quantized frequencies is small. Therefore, the quantization error and the PSNR are also high. Conversely, for high-frequency images, the number of valid frequencies increases, and the quantization error increases. Thus, PSNR is related to the characteristics of the image.

Therefore, the optimal image restoration effect can be obtained by matching the image degradation with the image restoration strength. In the restoration of the coded image, it is difficult to determine the optimal CycleGAN network because the input QP and training QP do not match. In this article, the network is trained to emphasize the improvement of the image distortion part of the coding and decoding process. Consequently, the Peak Signal-to-Noise Ratio (PSNR) value is chosen as a criterion for image segmentation. The PSNR is a measure of image quality and is widely used in the field of image processing, such as image compression, denoising, and super-resolution. The PSNR value is used to evaluate the quality of an image by comparing the difference between the original image and the image after processing (e.g., compression, enhancement, etc.). A higher PSNR value indicates a superior image quality with minimal distortion. The PSNR is defined in terms of the Mean Squared Error (MSE), which is calculated using the following formula:

$$MSE = \frac{1}{mn} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} |I(i, j) - K(i, j)|^2, \quad (6)$$

where I and K represent the original image and the processed image, respectively, and m and n are the number of rows and columns of the image, respectively. The PSNR is calculated in accordance with the following formula:

$$PSNR = 10 \times \log_{10} \left(\frac{MAX_I^2}{MSE} \right) \quad (7)$$

The objective of this study is to classify the datasets based on the PSNR values and to train a modified CycleGAN on each dataset within the specified PSNR range. This approach guarantees the uniformity of dataset degradation and facilitates the reduction process to align with the image features. Specifically, during the training phase, as illustrated in Figure 7, the datasets are classified at intervals of 2 dB, with each classified dataset subsequently employed for training purposes. Given that the coded images comprise numerous flat regions, it was deemed appropriate to set a larger image block size, thus facilitating the capture of the aforementioned features. The datasets are then cropped into image blocks of 128×128 pixels. In the inference phase, the PSNR values of each original and encoded image block are calculated, and the corresponding network weights are selected based on the PSNR values of each image block. To illustrate, if the PSNR value of the input–output image block is 35 dB, the network weights corresponding to the ‘34–37 dB PSNR range’ will be selected. The network will undergo training based on different PSNR value ranges of image blocks, with each range being trained individually. Additionally, the network weights will be modified dynamically during the training process to allow for repeated iterations and the achievement of optimal results. Once the aforementioned mechanism has been applied to all image blocks, the restored image blocks are stitched together to construct a completely restored image. By applying the network training results to each image block corresponding to its PSNR, the optimal reduction method can be applied to each part of the image.

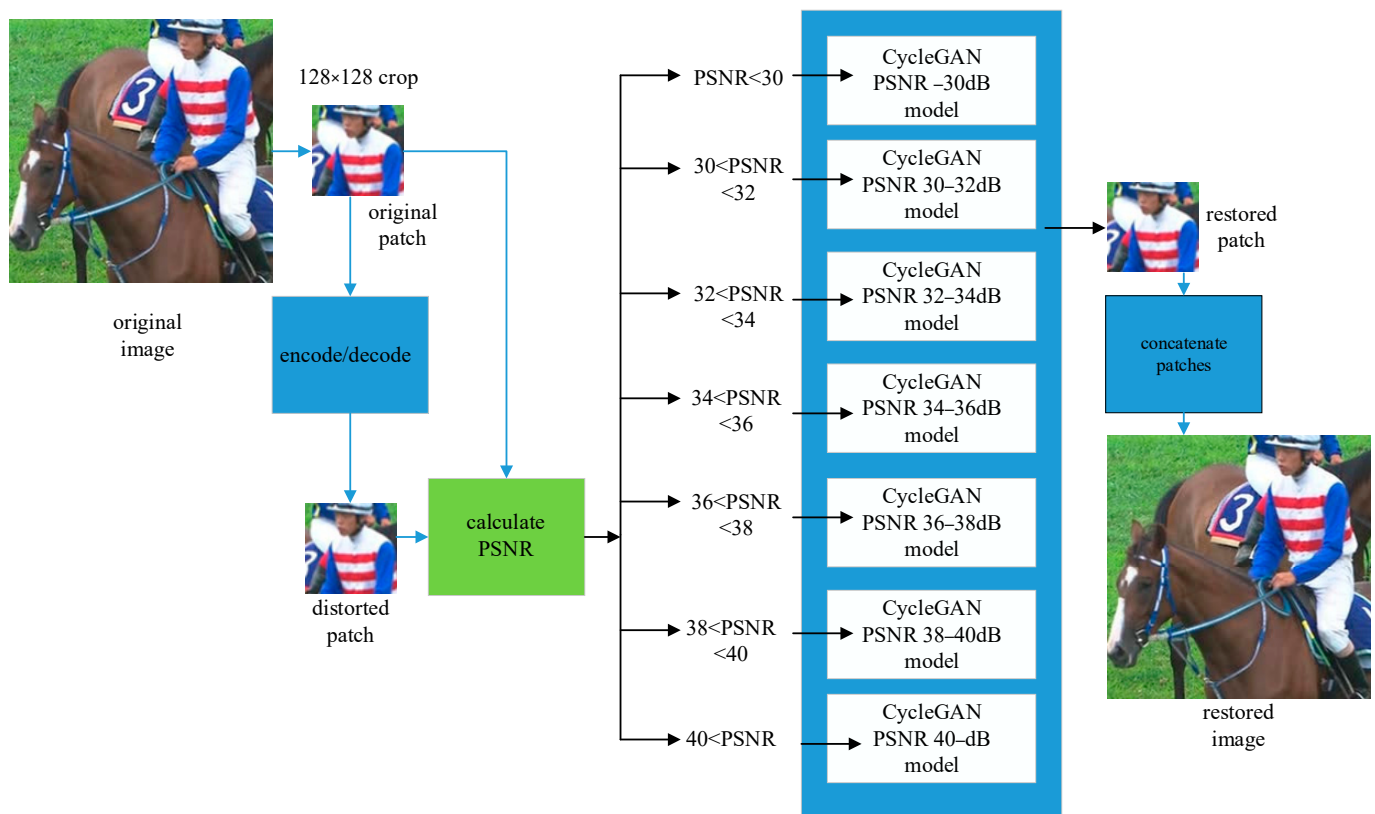


Figure 7. Training and inference adapted to PSNR.

3. Results and Discussion

3.1. Dataset Preparation

The dataset employed in this study was derived from the YUV Trace Dataset, which was made available by Arizona State University. It encompassed a multitude of video sequences in YUV format [16], offering a valuable resource for research in the domain of video processing and compression. The dataset comprised a diverse range of content, including people, scenery, buildings, creatures, and more common images of human landscapes. In this paper, eight video sequences—Akiyo, Bus, Bridge, Foreman, Flower, Container, Tempete, and Basketball—were selected for analysis. Each frame was extracted, and frames that were too similar to each other were removed to ensure the integrity of the dataset. A total of 1000 training images were selected, and the training set, validation set, and test set were divided according to a ratio of 8:1:1, resulting in 800 training sets, 100 validation sets, and 100 test sets, respectively. The training set was trained using the training method described in Section 2, and the test set was predicted based on the training results. The present study will initially compare the network improvement before and after the improvement. This will be followed by an analysis of the ablation experiment. Thereafter, the improvement effect under different QP values will be explored. Finally, this method will be compared with two other common network methods (ESRGAN and DRRN).

3.2. Experimental Environment and Evaluation Indicators

The hardware configuration and environment utilized for the experiments presented in this paper are detailed in Table 1.

In order to provide a comprehensive evaluation of the performance of the enhanced CycleGAN model, we employ not only PSNR values as metrics for assessment but also LPIPS values. This enables a comparison of the impact of human visual perception. LPIPS assesses the degree of visual similarity between two images by simulating the human visual perception process. This approach focuses on evaluating the visual similarity of the images rather than solely on pixel-level similarity. LPIPS employs a trained convolutional

neural network to discern image features that encapsulate both the visual content and structural aspects of an image. Consequently, it is a prevalent tool for assessing image quality, particularly in contexts where comparing the quality of images before and after operations such as image restoration, compression, or enhancement is of interest. The LPIPS values typically range from 0 to positive infinity. A value of 0 indicates that the two images are visually identical, whereas a positive value indicates the presence of a visual difference. The magnitude of the difference is reflected in the value, with higher values denoting a more pronounced discrepancy.

Table 1. Experimental configurations.

Configuration Names	Specific Environment
Operating System	Windows 10
CPU	i7-14900 kf
GPU	RTX3090 × 2
RAM Size	32 GB
Memory Size	24 GB × 2
Deep Learning Frameworks	Pytorch1.12
Cuda Version	Cuda11.4
Batch Size	32
Num Works	4

3.3. Comparison of Network Effect Before and After Improvement

A QP value of 32 was selected for comparison of the pre-improved CycleGAN and the improved CycleGAN network in restoring the distortion of the individual YUV video sequences mentioned in Section 3.1. The resulting PSNR values are presented below.

As illustrated in Table 2, the PSNR values of the enhanced CycleGAN network exhibit a notable improvement of an average of 5.0944% for the same coded QP values across the tested video sequences, which were initially encoded and decoded using the unimproved and enhanced CycleGAN networks, respectively. To assess the restoration performance, we utilized the LPIPS value as an evaluation metric, resulting in the data presented in Table 3.

Table 2. Comparison of PSNR values before and after improvement.

Video Sequences	Pre-Improved	Improved
Akiyo	29.4268	30.9709
Bus	32.9634	34.6446
Bridge	27.9703	29.3993
Foreman	33.0887	34.7747
Flower	31.3318	32.9228
Container	30.0143	31.5415
Tempete	26.9252	28.2482
Basketball	29.9703	31.5014

As illustrated in Table 3, the LPIPS values of the network for video restoration exhibited a mean reduction of an average of 8.2353% following the implementation of enhancements. This observation suggests that the upgraded network displays a discernible advantage in terms of human visual perception.

Table 3. Comparison of LPIPS values before and after improvement.

Video Sequences	Pre-Improved	Improved
Akiyo	0.2822	0.2527
Bus	0.2796	0.2504
Bridge	0.2845	0.2548
Foreman	0.2866	0.2567
Flower	0.2891	0.2589
Container	0.2818	0.2524
Tempete	0.2758	0.2470
Basketball	0.2839	0.2542

3.4. Ablation Experiments

This article presents the results of ablation experiments conducted to analyze the comparative effects of several factors on the image restoration process. The factors in question are the effects of increasing the attention mechanism, adjusting the loss function and changing the training scheme. In the experiment, the video sequence was initially processed by CycleGAN following the basic VVC codec and then compared with the results obtained after various processing methods. As can be seen from Table 4, the PSNR value of the video sequence before and after the initial processing is 27.6238, and the LPIPS value is 0.2946. The first processing method was to add the attention mechanism on the basis of the original network structure. The PSNR value increased by 2%, while the LPIPS value decreased by 3.02%, representing improvements in both cases. Method 2 involved the use of an enhanced loss function on the foundation of Method 1, resulting in a negligible change in the PSNR value and a further decrease in the LPIPS value. The PSNR decreased by 0.46%, while the LPIPS value decreased by 3.18%, indicating that the human eye perception was still enhanced. Processing Method 3 altered the training method based on processing Method 2, resulting in an increase of 7.38% in the PSNR and a decrease of 4.59% in the LPIPS value. Both metrics demonstrated a notable improvement.

Table 4. Comparison of ablation results.

Processing Method	Attention Mechanism	Loss Function	Training Method	PSNR	LPIPS
CycleGAN	✗	✗	✗	27.6238	0.2946
Method1	✓	✗	✗	28.1762	0.2857
Method2	✓	✓	✗	28.0459	0.2766
Method3	✓	✓	✓	30.1147	0.2639

3.5. Comparison of Results at Various QP Values

In order to verify the improvement effect of the improved method under the influence of different QP values, additional inference experiments with QP values of 37 and 35 were selected, and the results obtained are as follows in Tables 5 and 6.

According to the results in the table above, the improved method improves the PSNR by an average of 6.4462% and reduces the LPIPS value by an average of 6.4725% at QP = 37. At QP = 35, the improved method improves the PSNR by an average of 10.9875% and reduces the LPIPS value by an average of 10.7312%. Combined with the results at QP = 32 in Section 3.3, it shows that for image restoration after processing with different QP values, the improved method can achieve good results.

Table 5. Comparison of PSNR/LPIPS values (QP = 37).

Video Sequences	Pre-Improved	Improved
Akiyo	21.1844/0.3104	23.8498/0.2768
Bus	26.6978/0.3155	28.4123/0.3130
Bridge	21.2134/0.3129	21.7755/0.2804
Foreman	28.7527/0.3205	29.6062/0.3115
Flower	26.3201/0.3250	26.9888/0.3120
Container	23.4114/0.3109	25.5351/0.2776
Tempete	20.1939/0.2996	21.4672/0.2715
Basketball	25.4732/0.3285	27.7612/0.3180

Table 6. Comparison of PSNR/LPIPS values (QP = 35).

Video Sequences	Pre-Improved	Improved
Akiyo	25.0139/0.2971	28.4538/0.2643
Bus	29.6670/0.2958	30.4636/0.2654
Bridge	23.7735/0.2998	27.9192/0.2693
Foreman	29.7798/0.3018	31.2972/0.2683
Flower	26.5820/0.3051	30.6072/0.2729
Container	27.0128/0.2969	28.7121/0.2669
Tempete	22.8407/0.2918	24.5577/0.2585
Basketball	26.9732/0.2990	29.6312/0.2667

3.6. Comparison with Other Algorithms

In this section, the case where the QP value is uniformly 32 is selected for comparison of the effect of the improved method with the other two common GAN types of network utilization (ESRGAN and DRRN). The results are presented below in Table 7.

Table 7. Comparison of PSNR/LPIPS values (with ESRGAN and DRRN).

Video Sequences	Improved	ESRGAN	DRRN
Akiyo	30.9709/0.2527	28.3936/0.2605	26.9139/0.2865
Bus	34.6446/0.2504	31.7227/0.2592	29.9459/0.2877
Bridge	29.3993/0.2548	27.2089/0.2625	25.8484/0.2887
Foreman	34.7747/0.2567	32.3110/0.2654	30.5219/0.2945
Flower	32.9228/0.2589	29.9498/0.2670	28.4523/0.2815
Container	31.5415/0.2524	28.9182/0.2612	27.3715/0.2873
Tempete	28.2482/0.2470	26.5720/0.2549	25.1924/0.2831
Basketball	31.5014/0.2542	29.2864/0.2630	27.6580/0.2893

As evidenced by the data presented in the preceding table, the PSNR value of the enhanced methodology has exhibited an average increase of 8.3816% in comparison to ESRGAN and an average increase of 14.4% in comparison to DRRN. The LPIPS value has been reduced by an average of 4.8817% as compared to ESRGAN and by an average of 11.9500% as compared to DRRN. In comparison to DRRN, the LPIPS value has been reduced by an average of 654% as compared to ESRGAN.

4. Conclusions

This article proposes a method based on an enhanced CycleGAN model for the reduction in image distortion during VCC video codec. The method incorporates a lightweight attention mechanism into the original GAN generator component and improves the loss function used in training based on human eye perception. Furthermore, the training method is modified from the original direct training of the image dataset to one that involves the slicing and dicing of each piece of original data into small image blocks, which are then trained based on the PSNR parameters of the image blocks. Training is conducted in accordance with the PSNR parameters of the image blocks. The experimental results demonstrate that the average PSNR of the enhanced video image is 30.1147, and the LPIPS value is 0.2639, which effectively addresses the issue of image quality degradation during transmission. In the future, the enhanced network model will be implemented on high-performance video transmission infrastructure and integrated with the VCC decoding protocol to enhance the efficiency of video transmission while maintaining superior image quality, thereby better serving the needs of viewers and consumers.

Author Contributions: Conceptualization, W.D. and Z.Y.; methodology, W.D.; software, W.D.; validation, W.D. and Z.Y.; formal analysis, W.D.; investigation, W.D.; resources, Z.Y.; data curation, W.D.; writing—original draft preparation, W.D.; writing—review and editing, Z.Y.; visualization, W.D.; supervision, Z.Y.; project administration, Z.Y.; funding acquisition, Z.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the article; further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Uhrina, M.; Sevcik, L.; Bienik, J.; Smatanova, L. Performance Comparison of VVC, AV1, HEVC, and AVC for High Resolutions. *Electronics* **2024**, *13*, 953. [\[CrossRef\]](#)
2. Guo, J.; Li, S.; Zhu, J.; Li, X.; Sun, B.; Feng, W. Adaptive Transmission Strategy for Non-Uniform Coding of 360° Videos. *Electronics* **2024**, *13*, 3266. [\[CrossRef\]](#)
3. Chen, L.; Cheng, B.; Zhu, H.; Qin, H.; Deng, L.; Luo, L. Fast Versatile Video Coding (VVC) Intra Coding for Power-Constrained Applications. *Electronics* **2024**, *13*, 2150. [\[CrossRef\]](#)
4. HoangVan, X. Adaptive Quantization Parameter Estimation for HEVC Based Surveillance Scalable Video Coding. *Electronics* **2020**, *9*, 915. [\[CrossRef\]](#)
5. Dong, C.; Loy, C.C.; He, K.; Tang, X. Image super-resolution using deep convolutional networks. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 184–199.
6. Kim, J.; Lee, J.K.; Lee, K.M. Accurate image super-resolution using very deep convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 1646–1654.
7. Park, J.; Lee, J.; Sim, D. Low-complexity CNN with 1D and 2D filters for super-resolution. *J. Real-Time Image Process.* **2020**, *17*, 2065–2076. [\[CrossRef\]](#)
8. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: New York, NY, USA, 2014; Volume 27, pp. 2672–2680.
9. Zhu, J.-Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2242–2251.
10. Zhang, R.; Isola, P.; Efros, A.A.; Shechtman, E.; Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018.
11. Swain, S.; Tripathy, A.K. Automatic detection of potholes using VGG-16 pre-trained network and Convolutional Neural Network. *Heliyon* **2024**, *10*, e30957. [\[CrossRef\]](#) [\[PubMed\]](#)

12. Hemand, E.P.; Mohandass, G.; Shajin, F.H.; Kirubakaran, D. AlexNet-based deep convolutional neural network optimized with group teaching optimization algorithm (GTOA) for paediatric bone age assessment from hand X-ray images. *Imaging Sci. J.* **2024**, *72*, 336–348. [[CrossRef](#)]
13. Fardo, F.A.; Conforto, V.H.; de Oliveira, F.C.; Rodrigues, P.S. *A Formal Evaluation of PSNR as Quality Measurement Parameter for Image Segmentation Algorithms*; CoRR: Los Angeles, CA, USA, 2016.
14. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [[CrossRef](#)] [[PubMed](#)]
15. Zhang, L.; Zhang, L.; Mou, X.; Zhang, D. FSIM: A feature similarity index for image quality assessment. *IEEE Trans. Image Process.* **2011**, *20*, 2378–2386. [[CrossRef](#)] [[PubMed](#)]
16. Wen, X.; Pan, Z.; Hu, Y.; Liu, J. Generative Adversarial Learning in YUV Color Space for Thin Cloud Removal on Satellite Imagery. *Remote Sens.* **2021**, *13*, 1079. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.