IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 25 June 2023, accepted 19 July 2023, date of publication 2 August 2023, date of current version 11 August 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3301145

RESEARCH ARTICLE

WCDANN: A Lightweight CNN Post-Processing Filter for VVC-Based Video Compression

HAO ZHANG[1], CHEOLKON JUNG[1], (Member, IEEE), DAN ZOU[2], AND MING LI[2]
[1]School of Electronic Engineering, Xidian University, Xi'an 710071, China
[2]Guangdong OPPO Mobile Telecommunications Corporation, Dongguan 523860, China

Corresponding author: Cheolkon Jung (zhengzk@xidian.edu.cn)

This work was supported by the National Natural Science Foundation of China under Grant 62111540272.

ABSTRACT In this paper, we propose a weakly connected dense attention neural network for compression artifact removal, called WCDANN. WCDANN is a convolutional neural network (CNN)-based post-processing filter to enhance the quality of versatile video coding (VVC)-decoded videos without requiring any codec changes. WCDANN consists of several weakly connected dense attention blocks (WCDABs) based on residual learning, which takes the compressed video after codecs as the input. We use depthwise separable convolution for WCDANN as the basic convolution unit to generate a lightweight model. Moreover, we introduce attention mechanisms into the proposed filter to capture important features. Experimental results show that WCDANN achieves good performance in Bjøntegaard Delta Bit Rate (BD-BR). Compared with VTM-11.0-NNVC anchor, WCDANN achieves average 2.81%, 4.12% and 3.81% BD-rate reductions for Y channel on A1, A2, B, C, D and E classes in RA, AI and LDP configurations, respectively.

INDEX TERMS Video compression, attention, convolutional neural network, depthwise separable convolution, in-loop filter, post-processing.

I. INTRODUCTION

With recent advances in video communications, video-related applications are increasing day by day. Due to limited network bandwidth, efficient video compression technology is essential to transmit massive videos. Versatile Video Coding (VVC) [1] established by the Joint Video Experts Team (JVET) in July 2020 achieves better performance than HEVC [2]. VVC shows great performance by integrating numerous advanced features and functions for high spatial resolution, high dynamic range and 360° video formats, achieving up to 30% BD rate savings at the same quality as HEVC performance.

In VVC, the traditional block-based hybrid video coding architecture has been maintained, while many new coding tools have been introduced. Each individual encoding tool has different coding efficiency. As shown in Fig. 1, the framework includes transform, quantization, entropy coding, intra prediction, inter prediction, and loop filter. Among them, the transform unit transforms a video frame from the time domain to the frequency domain and concentrates energy in the low-frequency regions. The quantization module is used to reduce the dynamic range of the image, which is also the root cause of video coding distortion. Entropy coding is a technique utilized in the encoding process to transform data into binary streams, thereby facilitating efficient storage and transmission of data. Intra and inter prediction techniques are employed to remove spatial and temporal redundancy, respectively. To enhance the quality of video frames and optimize compression efficiency, a loop filter module is applied during the process of video coding, which includes deblocking filter (DBF), sample adaptive offset (SAO) and adaptive loop filter (ALF). Among them, DBF and SAO are two filters designed to reduce artifacts caused by the encoding process. DBF focuses on visual artifacts at block boundaries, while SAO complementarily reduces artifacts that may arise from quantization of transform coefficients within blocks. ALF is an adaptive filter for the reconstructed signal, reducing the mean square error (MSE) between the original and reconstructed samples based on Wiener-based adaptive filtering [3], [4].

The associate editor coordinating the review of this manuscript and approving it for publication was Victor Sanchez.

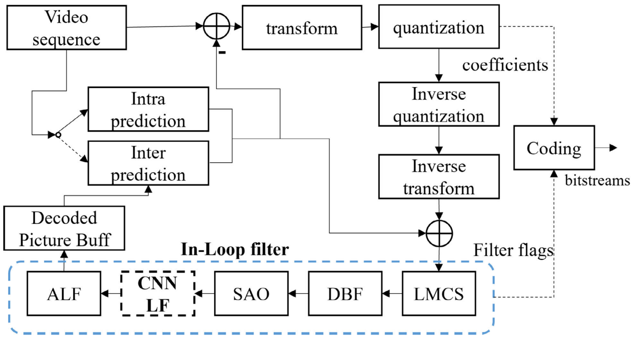


FIGURE 1. Illustration of CNN-based in-loop filter (CNNLF).

Although VVC maintains a high-quality compressed video through additional encoding functions, it is still in the process of continuous improvement, and researchers are constantly proposing new technologies to improve the coding performance of VVC. Reference [5] explored the intra-coding mode to improve coding efficiency, which is based on intra-sub partition coding modes to extend line-based intra-prediction modes. They introduced an adaptive sub-partition mechanism instead of a fixed number of sub-partitions, and achieved significant bitrate savings. At the same time, although the loop filter greatly suppresses compression artifacts, it is handcrafted and developed based on signal processing theory, assuming stationary signals. However, natural video sequences are usually non-stationary, thus the performance improvement by the loop filter is limited. Therefore, VVC still has a lot of room for improvement. Reference [6] proposed a CNN-based fast Coding Unit (CU) partitioning method for intra coding to reduce the coding complexity, which accelerates CU partition through predicting the partition modes with texture information and terminating redundant modes in advance. Corresponding classifiers are designed for different CU sizes to improve prediction accuracy.

With the advent of deep learning, many image and video quality enhancement methods based on convolutional neural networks (CNNs) have been proposed. Some CNN-based image enhancement methods [7], [8], [9], [10], [11], [12] were proposed to remove artifacts and blocking effects in JPEG compressed images, while other methods [13], [14], [15], [16], [17], [18] were proposed to improve the quality of HEVC compressed videos. These methods have achieved outstanding performance, demonstrating the effectiveness of CNNs in image quality enhancement. Recently, some studies are conducted to improve the quality of VVC compressed images and videos. These works are mainly divided into two categories: CNN-based in-loop filters (CNNLFs) [8], [19], [20], [21], [22] and CNN-based post-processing filters (CNNPPs) [23], [24], [25], [26], [27], [28]. CNNLFs are embedded in the VVC loop by inserting the middle of the loop filter components or replacing several loop filter components, as illustrated in Fig. 1. CNNPPs are located after the video decoding end, and directly enhance the output video of the decoding end to reduce visual artifacts and their pipeline is shown in Fig. 2. Since existing networks are usually complex and the amount of parameters is quite large, it is required to investigate a lightweight network for VVC.

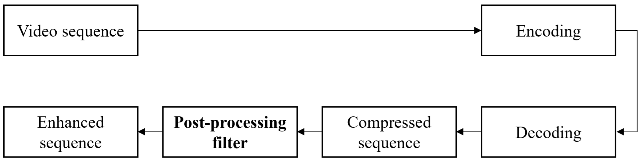


FIGURE 2. Pipeline of CNN-based post-processing filter (CNNPP).

In this paper, we propose a CNN post-processing filter based on depthwise separable convolution [29] and attention mechanism, named WCDANN. WCDANN aims to learn effective residual information and improve the quality of the input compressed image. WCDANN is composed of two basic modules, i.e. weakly connected dense attention block (WCDAB) and residual attention block (RAB) to extract important residual features from the input image. WCDANN includes residual connections in each module to promote the circulation of residual information and uses two attention modules of channel attention block (CAB) and channel spatial attention block (CSAB) to enhance important residual features in the output of RAB and WCDAB, respectively. We use depthwise separable convolution as the basic convolution unit to greatly reduce the amount of model parameters and generate a lightweight model. Furthermore, we train dedicated models for different quantization parameters (QPs). Finally, we compare the performance of standard convolution and depthwise separable convolution and demonstrate the effectiveness of depthwise separable convolution. Besides, we compare different embedding schemes and choose the best embedding method to evaluate the performance of the proposed filter as a loop filter in VVC Test Model (VTM).

Compared with existing methods, main contributions of this paper are summarized as follows:

- We propose a novel CNN based post-processing filter to enhance the quality of VVC-decoded video. The proposed CNN filter consists of several WCDABs and achieves good performance with relatively less parameters.
- We design a lightweight RAB as the basic unit of the proposed filter to extract features from a large receptive field and emphasize important channels from the extracted features.
- We propose a novel spatial attention module to obtain accurate spatial attention maps. Moreover, we use spatial attention and channel attention modules in CSAB to obtain a channel-spatial joint attention map.
- We explore various embedding methods of the proposed CNN filter into VVC loop filter and determine the optimal embedding scheme, thus providing a reference for embedding a CNN filter in VVC.

The rest of this paper is organized as follows. Section II reviews the related work. Section III presents the proposed method, including loss function and network details. Section IV provides the experimental results with the ablation study. Finally, we draw conclusion of this paper and outline future work in Section V.

II. RELATED WORK

A. IMAGE QUALITY ENHANCEMENT

With the recent advances in multimedia technology and network bandwidth, images and videos are widely used as a digital media for communications. However, due to the compression, artifacts and distortions are inevitably introduced into them, which lead to a unpleasant visual experience. Therefore, quality enhancement for lossy images and videos has become a key research issue. Recently, deep learning has achieved good performance in many fields, and more and more researchers use deep learning to enhance image quality. Inspired by image super-resolution based on CNN (SRCNN) [30], Dong et al. [7] proposed an end-to-end network structure for JPEG compression artifact removal, named ARCNN, which added one convolution layer to realize feature enhancement. It is the first compression artifact removal work based on deep learning. Following ARCNN, Zhang et al. [8] proposed DnCNN for image denoising, and used it for compression artifact removal. Tai et al. [9] proposed a persistent memory network (MemNet) which was stacked by memory blocks consisting of a recursive unit and a gate unit to learn explicit persistent memories. Qi et al. [10] proposed a subband adaptive image deblocking network based on wavelet and CNN to remove artifacts and blockiness of JPEG compressed images. Xie et al. [11] proposed a weakly connected dense generative adversarial network (WCDGAN) for compression artifact removal of highly compressed images. WCDGAN combined mixed convolution, weakly connected dense block, and mixed attention to obtain a generator, and combined four different loss functions as the total loss function to train the network. They generated photo-realistic images with compression artifact removal. Liang et al. [12] proposed a strong baseline model SwinIR for image restoration based on the Swin Transformer [31]. They used several residual Swin Transformer blocks (RSTB) to extract deep features, and each of RSTB has several Swin Transformer layers along with a residual connection. SwinIR achieved good performance in multiple vision tasks including JPEG compression artifact removal.

B. VIDEO QUALITY ENHANCEMENT

Based on image quality enhancement, video quality enhancement has also been extensively studied. In July 2020, the Joint Video Experts Group (JVET) established Versatile Video Coding (VVC), which achieves better performance than HEVC. Nonetheless, since lossy compression inevitably introduces distortion, quality enhancement for VVC has attracted attention by researchers.

1) CNN BASED IN-LOOP FILTER FOR VVC

Chen et al. [19] proposed a residual dense convolutional neural network (DRN) based in-loop filter for VVC. They utilized a residual learning module, dense shortcuts, and bottleneck layers to solve the gradient vanishing problem, encourage features reuse and save computational resources, respectively. Huang et al. [20] proposed a novel variable CNN (VCNN) based in-loop filter for VVC, which effectively handled the compressed videos with different QPs and FTs via a single model. Zhang et al. [8] proposed video quality enhancement for VVC based on a wide-activated squeeze-and excitation deep convolutional neural network (WSE-DCNN). They replaced VVC conventional in-loop filtering to eliminate the compression artifacts. Li et al. [21] proposed a CNN-based filter to enhance the quality of VVC intra coded frames, which took auxiliary information such as partition and prediction as input. For chroma channels, auxiliary information further included luma channel. Their filter achieved the best performance among neural network-based in-loop filters at the 20th JVET meeting. Wang et al. [22] presented a neural network based in-loop filter to replace DBF and SAO. They trained two models for I and B slices, and achieved good coding performance. Zhao et al. [32] proposed a model selection based multi-scale CNN model for in-loop filtering in VVC, which filtered the luminance and chrominance components simultaneously in the coding loop to improve the quality of the reconstructed frames. They adopted a model selection strategy to select the best CTU level model in terms of R-D cost at the encoder. Kathariya et al. [33] proposed a CNN-based in-loop filter to suppress compression artifacts in VVC. They utilized CNN features from the DCT transformed input to extract high-frequency components and introduce long-range correlation into the spatial CNN features by multi-stage feature fusion.

2) CNN BASED POST-PROCESSING FILTER FOR VVC

Zhang et al. [23] presented a novel CNN based post-processing method for VVC in the random access (RA) configuration. To achieve the optimal performance, they trained the network on a large video dataset containing VVC compressed content at various spatial resolutions, for different QP groups. The network is applied to the decoder to improve the reconstruction quality of VVC. Ma et al. [24] proposed a novel CNN based network, named MFRNet, for post-processing and in-loop filtering in the context of video compression. MFRNet consists of four multi-level feature review residual dense blocks (MFRBs), and each MFRB extracts features from multiple convolutional layers using a multi-level residual learning structure with dense connections while reusing high-dimensional features from previous MFRBs. This structure not only improves the reuse of features in the network, but also extracts the flow of information between blocks. Bonnineau et al. [28] investigated a learning-based solution as a post-processing step to enhance the decoded VVC video quality. They used multitask learning

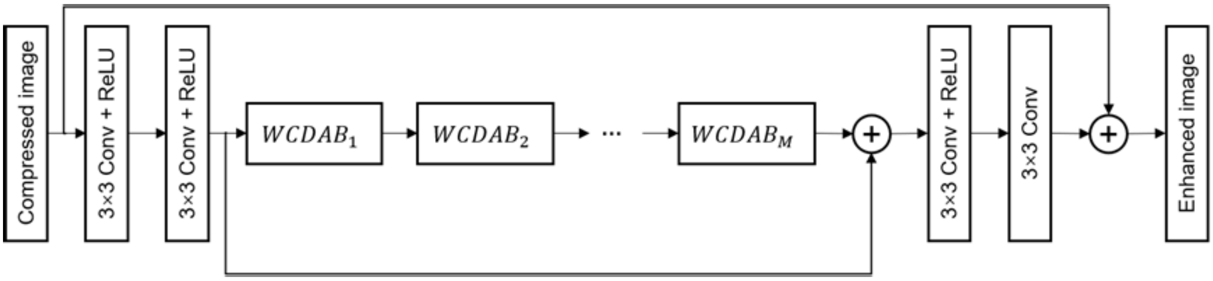


FIGURE 3. Network architecture of weakly connected dense attention neural network (WCDANN). WCDANN consists of M WCDABs. "3 × 3 Conv" denotes 3 × 3 common convolution layer. ReLU denotes Rectified Linear Unit for activation. Each Conv has 64 channels. $WCDAB_M$ denotes the M-th weakly connect dense attention block.

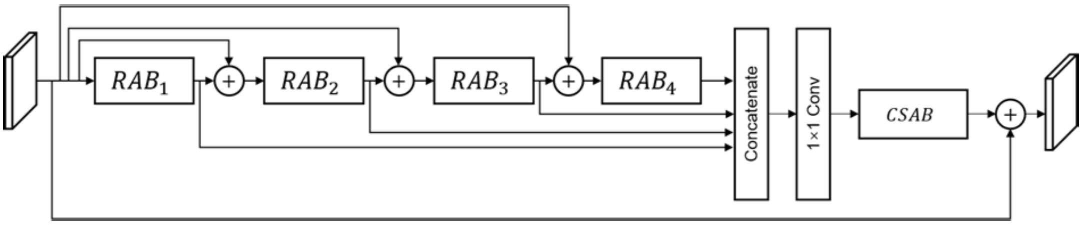


FIGURE 4. Network architecture of weakly connect dense attention block (WCDAB). WCDAB consists of M WCDABs. "1 × 1 Conv" denotes 1 × 1 common convolution layer. RAB_i denotes i-th residual attention block (i=1,···,4). CSAB: Channel-spatial joint attention block.

to perform both quality enhancement and super-resolution using a single shared network optimized for multiple degradation levels. Liu et al. [25] proposed a post-processing filter to improve the quality of VVC-decoded video based on a fusion network, named DFNN, which combined CNN and transformer [34] by channel-wise attention mechanism. Santamaria et al. [26] presented a CNN based content-adaptive post-processing filter to enhance VVC decoded video. This filter was content-adaptive, which was trained offline on general video sequences and later fine-tuned on the test video sequence. Subsequently, they enhanced the performance of the network and presented it at the 24th JVET meeting [27].

III. PROPOSED METHOD

A. OVERVIEW

The proposed weakly connected dense attention neural network (WCDANN) is mainly used to improve the quality of Y channel of VVC decoded videos. WCDANN is composed of three parts: Head, backbone, and reconstruction. Fig. 3 illustrates the network architecture of WCDANN. The head part consists of two convolutional layers, which are used to extract the shallow features of the input image. Each convolutional layer is followed by a ReLU activation function, and their kernel size is 3 × 3. The backbone part is the key component of the network and the most unique part of various networks. It is composed of M weakly connect dense attention blocks (WCDABs), and M is a hyperparameter. The reconstruction part is structurally symmetric with the head part, and the difference is that the ReLU activation function is removed after the last layer.

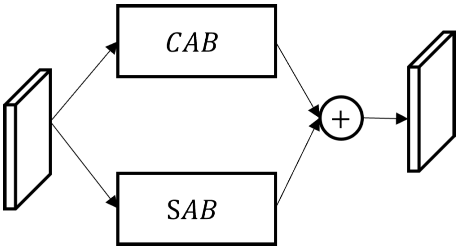


FIGURE 5. Network architecture of channel-spatial joint attention block (CSAB).

B. WEAKLY CONNECT DENSE ATTENTION BLOCK

The network architecture of the weakly connect dense attention block (WCDAB) is shown in Fig. 4. WCDAB consists of four residual attention blocks (RABs) and one channel-spatial joint attention block (CSAB). Some residual and concatenate connections are used in WCDAB to promote the circulation of residual information and prevent the gradient disappearance during training. We use one CSAB in the end of each WCDAB to emphasize spatial and channel-wise feature refinement (see the details of CSAB in Section III.C).

C. CHANNEL-SPATIAL JOINT ATTENTION BLOCK

The architecture of channel-spatial joint attention block (CSAB) is shown in Fig. 5. CSAB consists of two parallel branches: channel attention branch and spatial attention branch. The two branches consist of channel attention block (CAB) and spatial attention block (SAB), respectively. Through the two branches, the channel attention map and spatial attention map are obtained corresponding to the input

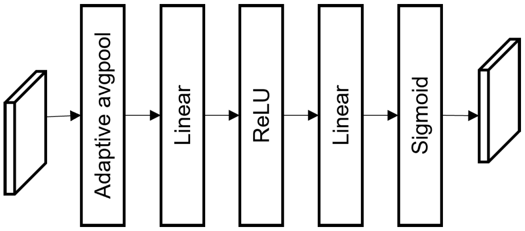


FIGURE 6. Network architecture of channel attention block (CAB).

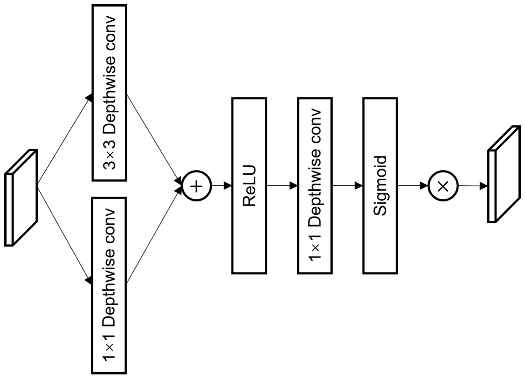


FIGURE 7. Network architecture of spatial attention block (SAB). 3 × 3 depthwise conv denotes 3 × 3 depthwise convolution layer. 1 × 1 depthwise conv denotes 1 × 1 depthwise convolution layer.

features, then they are fused through the addition operation to obtain the channel-spatial attention map. The channel attention module emphasizes the important residual feature by enhancing the features of useful channels and suppressing the features of useless channels. However, it is not accurate to enhance or suppress the entire channel only based on the importance of each channel. The unimportant channels also contain useful information in a certain feature space. After suppressing these channels, the useful information in the channels is suppressed. To make up for the useful information lost by the channel attention module, we use the spatial attention module to obtain the spatial attention map of each channel and add it to the channel attention map.

Next, we provide the specific architectures of CAB and SAB in Figs. 6 and 7, respectively. Our CAB follows a network architecture in the previous work [35]. It first extracts the weight of each channel through global average pooling, channel compression and expansion, then multiplies the extracted weight with the input feature map to generate the channel attention map. In SAB, for the input feature, we use parallel convolution kernels of different sizes to convolve it, then add the results of the two convolutions and use ReLU function to activate them. After that, we perform convolution and Sigmoid operation on the resultant feature map to get the spatial attention mask. Finally, we multiply the attention mask with the input to get the final spatial attention map.

The proposed SAB has two main differences from Compared with the existing spatial attention modules. First, all convolutions in SAB are depthwise convolutions. There are two reasons for using depthwise convolutions. On the one hand, spatial attention focuses on spatial information, thus ignoring the correlation between channels. The depthwise convolution is only spatially convolved on the feature map of each channel, and the relationship between the channels is not considered. Thus, from this viewpoint, it is reasonable to use the depthwise convolution for SAB. On the other hand, the proposed CSAB needs to be used at the end of each WCDAB, thus its parameters need to be affordable. In general, depthwise convolution requires quite less parameters than common convolution, thus using depthwise convolution can greatly reduce the amount of network parameters. Second, the number of channels of the attention mask obtained in SAB is the same as the input channel, which indicates that for each channel of the input SAB calculates the attention mask. Compared with the single-channel attention mask in existing methods, SAB can more accurately retain important spatial information in each channel.

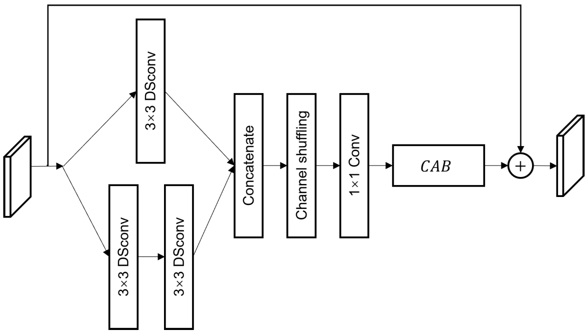


FIGURE 8. Network architecture of residual attention block (RAB). 3 × 3 DSconv denotes 3 × 3 depthwise separable convolution layer. Channel shuffle denotes channel shuffle operation. CAB: Channel attention block.

D. RESIDUAL ATTENTION BLOCK

The network architecture of the proposed residual attention block (RAB) is shown in Fig. 8. Each RAB has a dual-branch structure to increase the receptive field and capture multi-level information. All the convolution layers in these two branches are depthwise separable convolution, and their network architecture is shown in Fig. 9. The ReLU activation function can be used after only the pointwise convolutional layer, instead of the depthwise convolutional layer. We use a channel shuffle layer after two branches to fully integrate the features of different receptive fields at the channel level. Furthermore, we use the channel attention block (CAB) in the end of each RAB to emphasize channel-wise feature refinement.

E. LOSS FUNCTION

In the training stage, L1 loss and L2 loss are used to train WCDANN. Specifically, we use L1 loss in the first 40 epochs and L2 loss in the last 10 epochs. The total loss function is represented as follows:

$$L_{total} = \begin{cases} L_1(I_{out}, I_{gt}) & Epoch<40 \\ L_2(I_{out}, I_{gt}) & Epoch \geq 40 \end{cases} \tag{1}$$

where L_1 represents L1 loss, L_2 represents L2 loss. L1 loss and L2 loss is represented as follows:

$$L1(I_{out}, I_{gt}) = \frac{1}{N}\sum_{i=0}^{N}|I_{gt} - I_{out}| \tag{2}$$

$$L2(I_{out}, I_{gt}) = \frac{1}{N}\sum_{i=0}^{N}(I_{gt} - I_{out})^2 \tag{3}$$

where N is the number of training samples in each batch, I_{out} denotes the output of WCDANN, I_{gt} denotes the ground truth. The L1 loss enables the network to converge stably, while the L2 loss enables the network to converge further.

IV. EXPERIMENTAL RESULTS

First, we describe the experimental setup in detail. Then, we compare the proposed method with some latest works to demonstrate the advantages of the proposed method. Finally, we provide ablation studies on DSconv and embedding of the proposed filter into VVC Test Model (VTM).

A. EXPERIMENTAL SETTING

1) HARDWARE AND HYPERPARAMETERS

We implement WCDANN in PyTorch platform and perform training using a PC with NVIDIA GeForce GTX 3090 GPU. The total epochs are 50 and the batch size is set to 8. We empirically set the learning rate to 1e-4 for obtaining the optimal rate of convergence and stability.

2) TRAINING SET

We use DIV2K [36] and BVI-DVC [37] datasets to train WCDANN. All pictures are compressed using VTM 11.0-NNVC under the condition of QP = {22, 27, 32, 37, 42} to train the corresponding QP networks individually. All pictures are cropped into 128 × 128 patches in 125 steps to form the training set. We only use the Y channel of each image in the training set to train WCDANN.

3) NETWORK SETTING

In the training phase, we use two network architectures, and their main difference is the number of WCDABs. We refer to them as $WCDANN_{small}$ and $WCDANN_{big}$. $WCDANN_{small}$ includes 3 WCDABs (M=3), while $WCDANN_{big}$ includes 6 WCDABs (M=6). $WCDANN_{small}$ is used to enhance the quality of video using low QP compression (QP= 22, 27), and $WCDANN_{big}$ is used to enhance the quality of video with high QP compression (QP= 32,37,42). It is worth noting that we generate five models ($WCDANN_{QP_{22}}$, $WCDANN_{QP_{27}}$, $WCDANN_{QP_{32}}$, $WCDANN_{QP_{37}}$ and $WCDANN_{QP_{42}}$) during training according to five QPs (22, 27, 32, 37, 42) and fix the parameters of these five models until the training is complete. Since WCDANN is a post-processing filter used after the VTM decoder, there is no need to transmit the network parameters from the encoder to the decoder through bitstream or generate them in the decoder side.

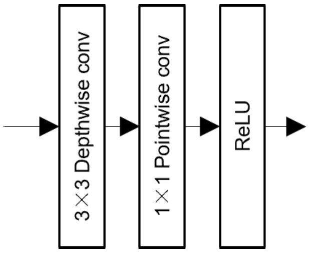


FIGURE 9. Network architecture of DSconv in our implementation.

TABLE 1. Performance comparison between the proposed method and VTM-11.0-NNVC anchor in RA configuration. MSIM: multi-scale structural similarity (MS-SSIM). Negative BD-rate values indicate coding gains.

Resolution	Class	Y-PSNR	Y-MSIM
3840x2160	A1	-2.23%	-1.78%
3840x2160	A2	-2.70%	-2.09%
1920x1080	B	-2.73%	-2.68%
832x480	C	-3.43%	-2.64%
416x240	D	-4.76%	-2.68%
1280x720	E	-	-
Overall		-2.81%	-2.37%

4) INFERENCE SETTING

In the inference phase, we use WCDANN according to base QPs [23] as follows:

$$L_{total} = \begin{cases} WCDANN_{QP_{22}} & QP_{base} \leq 24.5 \\ WCDANN_{QP_{27}} & 24.5<QP_{base} \leq 29.5 \\ WCDANN_{QP_{32}} & 29.5<QP_{base} \leq 34.5 \\ WCDANN_{QP_{37}} & 34.5<QP_{base} \leq 39.5 \\ WCDANN_{QP_{42}} & QP_{base}>39.5 \end{cases} \tag{4}$$

Compressed videos by different QPs are inferred on the corresponding models. We use the CTC test sequence (classes A1, A2, B, C, D and E) as the test set and Bjøntegaard Delta Bit Rate (BD-BR) [38] as the evaluation metric to evaluate the performance of the network. These test sequences are compressed using VTM-11.0-NNVC [39] at different QPs, and then the output compressed video is processed frame by frame using the trained model to obtain the enhanced video.

B. PERFORMANCE EVALUATION

1) OVERALL RD PERFORMANCE

The proposed CNN post-processing filter is evaluated in comparison with VTM-11.0-NNVC anchor according to the common test conditions (CTCs) defined in [40]. We perform the evaluation in the Random Access (RA), All Intra (AI) and Low Delay P (LDP) configurations. Experimental

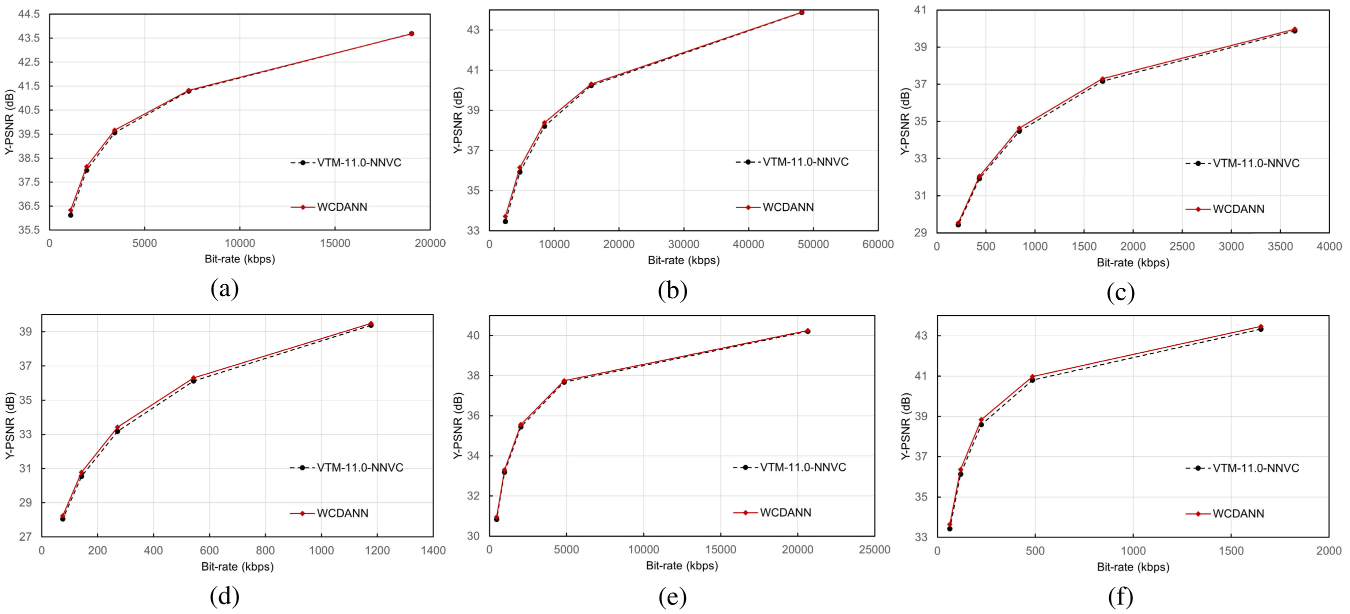


FIGURE 10. R-D curves by VTM-11.0-NNVC and WCDANN. (a) Class A1 (AI). (b) Class A2 (AI). (c) Class C (RA). (d) Class D (RA). (e) Class B (LDP). (f) Class E (LDP).

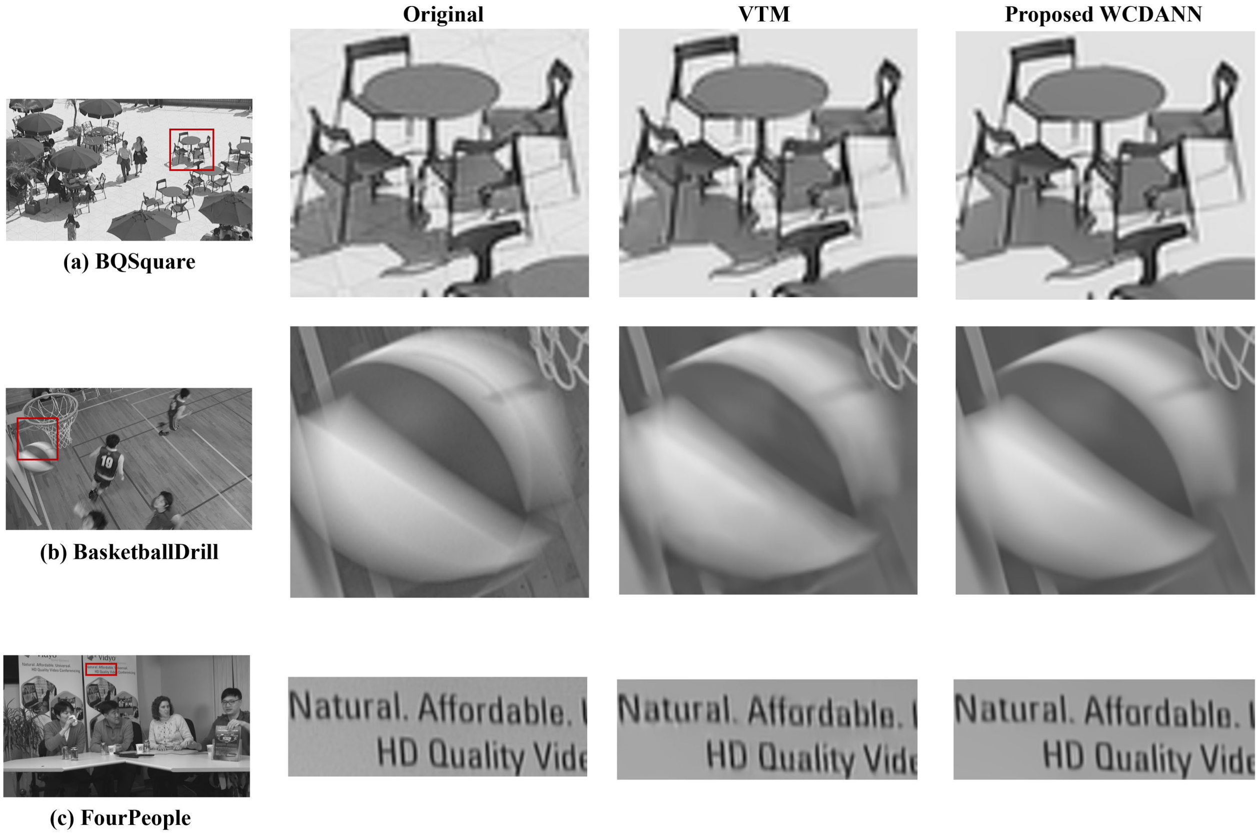
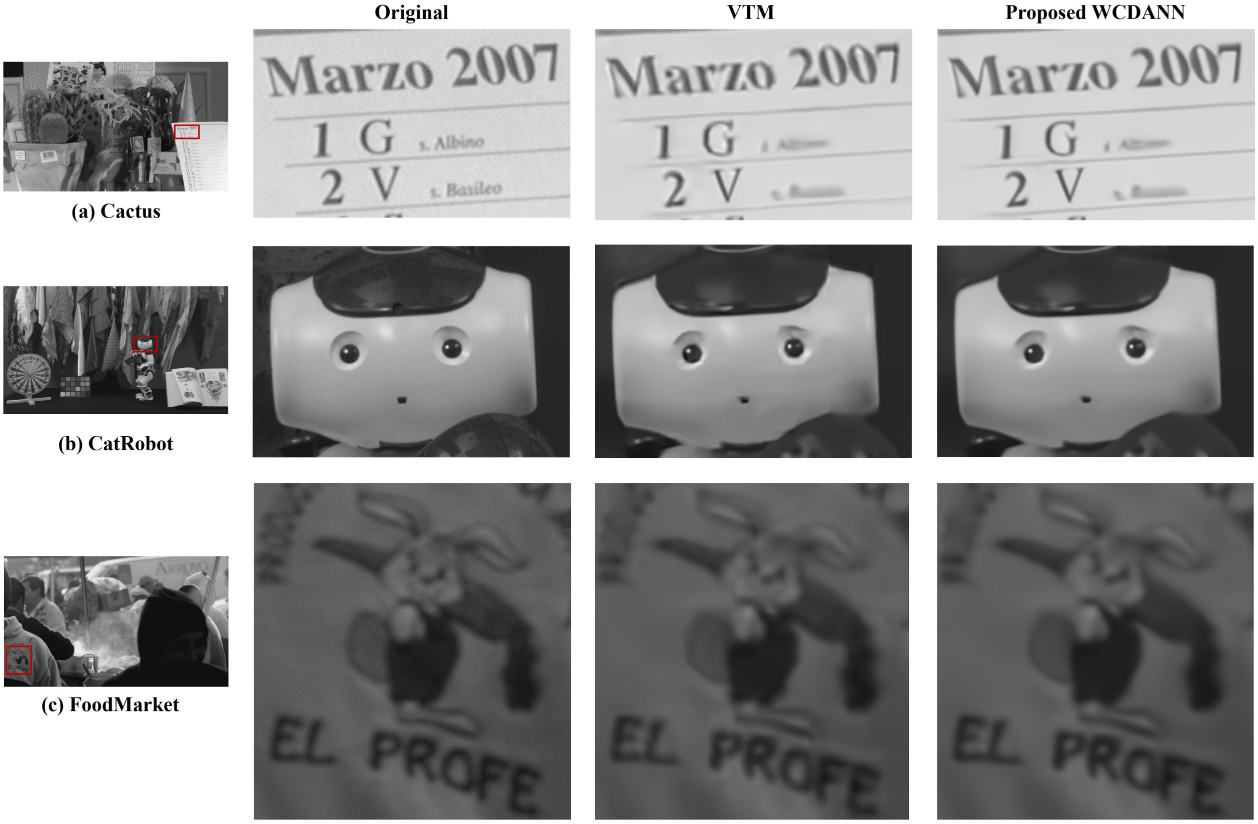


FIGURE 11. Comparison of visual quality between VTM-11.0-NNVC and WCDANN. (a) Class D-BQSquare (QP=37, RA). (b) Class C-BasketballDrill (QP=37, LDP). (c) Class E-FourPeople (QP=37, AI).

results demonstrate that the proposed CNN filter achieves average 2.81%, 4.12% and 3.81% BD-rate reductions over VTM 11.0-NNVC anchor for Y channel on A1, A2, B, C, D and E classes of the common test conditions (CTC) in RA, AI and LDP configurations, respectively. The test results under these three configurations (RA, AI and LDP) are shown



(a) Cactus

(b) CatRobot

(c) FoodMarket

FIGURE 12. Comparison of visual quality between VTM-11.0-NNVC and WCDANN. (a) Class B-Cactus (QP=42, RA). (b) Class A2-CatRobot (QP=42, LDP). (c) Class A1-FoodMarket (QP=42, AI).

TABLE 2. Performance comparison between the proposed method and VTM-11.0-NNVC anchor in AI configuration. MSIM: multi-scale structural similarity (MS-SSIM). Negative BD-rate values indicate coding gains.

Resolution	Class	Y-PSNR	Y-MSIM
3840x2160	A1	-2.98%	-2.85%
3840x2160	A2	-3.19%	-3.71%
1920x1080	B	-3.48%	-3.68%
832x480	C	-4.58%	-3.99%
416x240	D	-4.97%	-3.70%
1280x720	E	-6.66%	-7.85%
Overall		-4.12%	-4.31%

TABLE 3. Performance comparison between the proposed method and VTM-11.0-NNVC anchor in LDP configuration. MSIM: multi-scale structural similarity (MS-SSIM). Negative BD-rate values indicate coding gains.

Resolution	Class	Y-PSNR	Y-MSIM
3840x2160	A1	-2.53%	-2.75%
3840x2160	A2	-2.92%	-2.94%
1920x1080	B	-3.49%	-3.94%
832x480	C	-3.95%	-3.12%
416x240	D	-4.72%	-3.17%
1280x720	E	-6.33%	-7.30%
Overall		-3.81%	-3.95%

in Tables 1, 2, and 3, respectively. It should be noted that since CTC does not recommend testing video sequences in Class E in RA configuration, the RA test results of Class E are not given in the results. The rate-distortion (R-D) curves of some classes in some configurations are shown in Fig. 10. To show the RD curves of the proposed method in all sequences, we average the bitrate and PSNR on the test sequences in each class, and use the mean value to plot the RD curve for each class. It can be observed from the R-D curves that the proposed post-processing filter achieves better performance than VTM-11.0-NNVC anchor under multiple QPs and configurations.

2) COMPARISON WITH OTHER POST-PROCESSING FILTERS

To verify the performance of WCDANN, we compare it with two other CNN-based post-processing filters in RA configurations. These two methods, denoted JVET-X0111 [27] and JVET-Z0101 [25], have been presented at the 24th JVET meeting in October 2021 and the 26th JVET meeting in April 2022, respectively. The results are shown in Table 4. **Bold** indicates the best performance index obtained on each

TABLE 4. BD-rate comparison among the proposed method and two JVET contributions [25], [27] over VTM-11.0-NNVC anchor in RA configuration. MSIM: multi-scale structural similarity (MS-SSIM).

Method	Proposed		JVET-X0111		JVET-Z0101	
Metric	Y-PSNR	Y-MSIM	Y-PSNR	Y-MSIM	Y-PSNR	Y-MSIM
Class	BD-rate	BD-rate	BD-rate	BD-rate	BD-rate	BD-rate
A1	**-2.23**%	-1.78%	-2.06%	**-3.86**%	-0.80%	-1.02%
A2	**-2.70**%	**-2.09**%	-1.52%	-1.16%	-1.54%	-1.45%
B	-2.73%	**-2.68**%	**-3.10**%	-1.94%	0.02%	-0.84%
C	**-3.43**%	**-2.64**%	-3.34%	-1.60%	-1.91%	-1.40%
D	**-4.76**%	**-2.68**%	-3.92%	-1.15%	-4.09%	-1.99%
E	-	-	-	-	-	-
Overall	**-2.81**%	**-2.37**%	-2.64%	-2.08%	-0.97%	-1.15%

class among the various methods. As shown in the table, compared to JVET-X0111, although WCDANN performs worse in MS-SSIM for Class A1 and PSNR for Class B, it performs well in most classes. Compared with JVET-Z0101, WCDANN performs better in all cases. Overall, it can be concluded that WCDANN outperforms the other two methods on PSNR and multi-scale structural similarity (MSIM).

Moreover, we perform BD-rate comparison of WCDANN with some related state-of-the-art (SOTA) works [41], [42] in AI configuration. The results are shown in Table 5. It can be observed that WCDANN achieves the best performance in BD-rate except A1 class in terms of Y-PSNR and Y-MSIM. Since the proposed post-processing filter is located after the VVC decoder, it does not affect the encoding and decoding time of the VVC codec. That is, the encoding and decoding time of the VVC codec is not changed by WCDANN.

3) VISUAL COMPARISON

We provide visual comparison of WCDANN with VTM-11.0-NNVC anchor under three configurations (RA, LDP, AI) at QP=37. For experiments, we use three lower resolution video sequences: BQSquare (416 × 240), BasketballDrill (832 × 480) and FourPeople(1280 × 720). The visual comparison results are shown in Fig. 11. As shown in the figure, WCDANN effectively reduces the artifacts in the VTM compressed image (see the edge of the table and chairs, the outline of the basketball, and the edge of the letters), making the object edges clearer and greatly improving the visual quality of the image. Then, we set QP to 42, and choose three high resolution video sequences (Cactus (1920 × 1080), CatRobot (3840 × 2160) and FoodMarket (3840 × 2160)) to valid the visual quality. As shown in Fig. 12, WCDANN effectively removes ringing artifacts at the object edges (edges of letters and patterns) and the banding artifacts at the flat region (the robot face). The results verify that WCDANN still achieves good performance at high resolution.

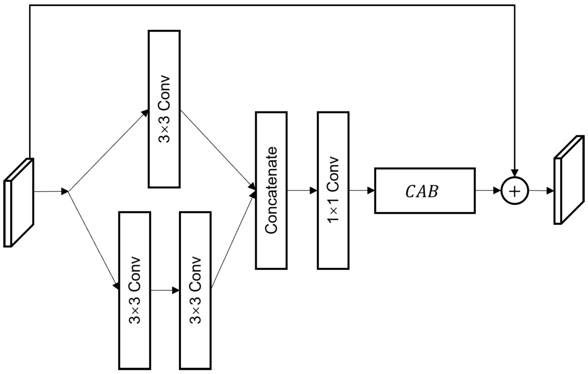


FIGURE 13. Network architecture of modified RAB.

C. ABLATION STUDY

The ablation studies include comparing the contributions of standard convolution and depthwise separable convolution to the performance as well as evaluating the performance of WCDANN as an in-loop filter in VTM.

1) DEPTHWISE SEPARABLE CONVOLUTION VS STANDARD CONVOLUTION

In this ablation experiment, all depthwise separable convolutions in WCDANN are replaced with standard convolutions and trained under the same setting. Specifically, we replace all depthwise separable convolutions in RAB while remaining the other structures unchanged. The modified RAB is shown in Fig. 13. For the convenience of distinction, we name the original WCDANN as WCDANN-DSConv, and the modified WCDANN as WCDANN-SConv. The test results are shown in Tables 6, 7 and 8. Compared with WCDANN-DSConv, WCDANN-SConv achieves average -0.20%, 0.42% and 0.12% BD-rate reductions for Y channel on each test classes, under RA, AI and LDP configurations, respectively. The results show that the depthwise separable convolution achieves similar performance to the standard convolution. Moreover, we test the parameters and flops of the two

TABLE 5. BD-rate comparison among the proposed method and two recently published works [41], [42] in AI configuration.

Method	Proposed in VTM-11.0-NNVC-1.0		Qi *et al.* [42] in VTM-11.0-NNVC-1.0		Cui *et al.* [43] in VTM-10.0	
Metric	Y-PSNR	Y-MSIM	Y-PSNR	Y-MSIM	Y-PSNR	Y-MSIM
Class	BD-rate	BD-rate	BD-rate	BD-rate	BD-rate	BD-rate
A1	-2.98%	**-2.85**%	-2.64%	-2.68%	**-2.99**%	-
A2	**-3.19**%	**-3.71**%	-3.19%	-3.71%	-2.74%	-
B	**-3.48**%	**-3.68**%	-3.17%	-2.93%	-3.36%	-
C	**-4.58**%	**-3.99**%	-4.31%	-3.39%	-3.98%	-
D	**-4.97**%	**-3.70**%	-4.58%	-3.18%	-4.76%	-
E	**-6.66**%	**-7.85**%	-5.51%	-6.29%	-5.42%	-
Overall	**-4.12**%	**-4.31**%	-3.90%	-3.27%	-3.88%	-

TABLE 6. Performance comparison between WCDANN-DSConv and WCDANN-SConv in RA configuration. MSIM: multi-scale structural similarity (MS-SSIM).

Method	WCDANN-DSConv		WCDANN-SConv	
Metric	Y-PSNR	Y-MSIM	Y-PSNR	Y-MSIM
Class	BD-rate	BD-rate	BD-rate	BD-rate
A1	-2.23%	-1.78%	-1.93%	-1.76%
A2	-2.70%	-2.09%	-2.38%	-1.74%
B	-2.73%	-2.68%	-2.45%	-2.32%
C	-3.43%	-2.64%	-3.47%	-2.35%
D	-4.76%	-2.68%	-4.75%	-2.56%
E	-	-	-	-
Overall	-2.81%	-2.37%	-2.61%	-2.10%

TABLE 7. Performance comparison between WCDANN-DSConv and WCDANN-SConv in AI configuration. MSIM: multi-scale structural similarity (MS-SSIM).

Method	WCDANN-DSConv		WCDANN-SConv	
Metric	Y-PSNR	Y-MSIM	Y-PSNR	Y-MSIM
Class	BD-rate	BD-rate	BD-rate	BD-rate
A1	-2.98%	-2.85%	-2.76%	-2.95%
A2	-3.19%	-3.71%	-3.50%	-4.10%
B	-3.48%	-3.68%	-3.81%	-4.06%
C	-4.58%	-3.99%	-5.77%	-4.66%
D	-4.97%	-3.70%	-5.43%	-4.18%
E	-6.66%	-7.85%	-6.95%	-7.81%
Overall	-4.12%	-4.31%	-4.54%	-4.64%

networks. The results are shown in Table 9. It can be observed that the performance of WCDANN-DSConv and WCDANN-SConv is very similar even though WCDANN-DSConv reduces almost three times the number of parameters.

TABLE 8. Performance comparison between WCDANN-DSConv and WCDANN-SConv in LDP configuration. MSIM: multi-scale structural similarity (MS-SSIM).

Method	WCDANN-DSConv		WCDANN-SConv	
Metric	Y-PSNR	Y-MSIM	Y-PSNR	Y-MSIM
Class	BD-rate	BD-rate	BD-rate	BD-rate
A1	-2.53%	-2.75%	-2.29%	-2.75%
A2	-2.92%	-2.94%	-3.05%	-3.02%
B	-3.49%	-3.94%	-3.74%	-3.64%
C	-3.95%	-3.12%	-4.45%	-3.54%
D	-4.72%	-3.17%	-4.92%	-3.54%
E	-6.33%	-7.30%	-6.07%	-6.79%
Overall	-3.81%	-3.95%	-3.93%	-3.89%

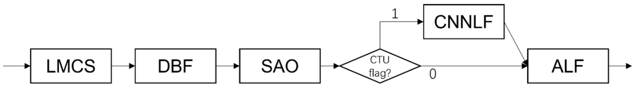


FIGURE 14. Pipeline of LDSCA. L: LMCS. D: DBF. S: SAO. C: CNNLF. A: ALF.

TABLE 9. Comparison of model complexity between WCDANN-DSConv and WCDANN-SConv.

Model	Parameter(M)		Flops(G)(128 × 128)	
WCDANN DSConv	$WCDANN_{small}$	0.43	$WCDANN_{small}$	6.55
	$WCDANN_{big}$	0.79	$WCDANN_{big}$	11.88
WCDANN SConv	$WCDANN_{small}$	1.59	$WCDANN_{small}$	25.50
	$WCDANN_{big}$	3.10	$WCDANN_{big}$	49.78

2) WCDANN FOR CHROMA COMPONENTS

For the compressed videos with the YUV format, the Y component is dominant, and its quality directly affects the visual quality of the decoded videos. Therefore, we first train WCDANN for the Y component to improve the quality of the decoded videos. To handle chroma components, we then

TABLE 10. Performance for YUV channels in RA configuration. MSIM: multi-scale structural similarity (MS-SSIM).

Metric	Y-PSNR	U-PSNR	V-PSNR	Y-MSIM	U-MSIM	V-MSIM
Class	BD-rate	BD-rate	BD-rate	BD-rate	BD-rate	BD-rate
A1	-2.23%	-3.73%	-3.94%	-1.78%	-4.17%	-4.77%
A2	-2.70%	-3.67%	-2.98%	-2.09%	-4.21%	-1.68%
B	-2.73%	-2.59%	-4.01%	-2.68%	-4.18%	-4.68%
C	-3.43%	-3.88%	-6.06%	-2.64%	-5.59%	-7.71%
D	-4.76%	-3.32%	-5.62%	-2.68%	-4.69%	-6.74%
E	-	-	-	-	-	-
Overall	-2.81%	-3.38%	-4.34%	-2.37%	-4.56%	-4.90%

TABLE 11. Performance for YUV channels in AI configuration. MSIM: multi-scale structural similarity (MS-SSIM).

Metric	Y-PSNR	U-PSNR	V-PSNR	Y-MSIM	U-MSIM	V-MSIM
Class	BD-rate	BD-rate	BD-rate	BD-rate	BD-rate	BD-rate
A1	-2.98%	-2.78%	-4.44%	-2.85%	-4.15%	-6.04%
A2	-3.19%	-4.24%	-4.45%	-3.71%	-4.86%	-3.55%
B	-3.48%	-4.16%	-5.67%	-3.68%	-5.56%	-6.45%
C	-4.58%	-4.66%	-7.30%	-3.99%	-6.38%	-8.90%
D	-4.97%	-4.37%	-7.21%	-3.70%	-6.22%	-9.30%
E	-6.66%	-6.60%	-9.27%	-7.85%	-7.38%	-10.09%
Overall	-4.12%	-4.46%	-6.22%	-4.31%	-5.69%	-7.05%

train WCDANN on the UV training set and obtain the results in chroma components. We set the number of WCDABs to 3 and use the UV component as the network input to train the network using the same training method as the Y component. We test the CTC results in AI and RA configurations under {Y, U, V} channels, the results are shown in Table 10 and Table 11. Experimental results demonstrate that the proposed WCDANN achieves average {2.81%, 3.38%, 4.34%} and {4.12%, 4.46%, 6.22%} BD-rate reductions on {Y, U, V} channels in RA and AI configurations, respectively.

3) EMBEDDING IN VTM

In this stage, we explore the performance of embedding WCDANN in VTM-11.0-NNVC as a CNN loop filter (CNNLF). It should be noted that the WCDANN is not retrained during this process, and all network coefficients are the same as the post-processing filter. To find the optimal embedding method, we conduct experiments on five embedding methods and compare the corresponding performance of each embedding method on C classes in RA configurations. According to the combination in the loop filter, these embedding methods are named as: LCDSA, LDCSA, LDSCA, LDSAC, and LCA, where L, C, D, S, A are LMCS, CNNLF, DBF, SAO, ALF, respectively. For the above five embedding methods, we select the filter according to the QP value of each frame in the same way as in the inference stage, and use CTU as the filtering unit to process each frame of video. We enable the slice level flag and the CTU flag, which means the proposed filter can be turned on/off at the CTU level and slice level. After processing a CTU or slice, the RD cost before and after processing is calculated to measure whether the slice level flag and CTU flag are enabled. It is obvious that turning off DBF and SAO for I slice can improve the performance, while for P and B slice, DBF and SAO need to be enabled. Therefore, we apply this rule to each embedding method. The comparison results among the embedding methods are shown in Table 12. As shown in Table 12, LDCSA has higher performance than LCDSA, which shows that DBF behind WCDANN degrades performance. The performances of LDCSA and LDSCA are similar, indicating that SAO is hardly activated during the process. Compared with LDSCA, the performance of LDSAC is degraded, which shows that ALF has a great effect on the performance improvement. Finally, we use WCDANN to replace DBF and SAO, i.e. LCA. The LCA does not perform well, which shows that WCDANN is not suitable for replacing DBF and SAO, and confirms that WCDANN is not suitable for processing the signal before DBF.

The optimal schemes are LDCSA and LDSCA. Since the post-processing filters are all enabled when we compress the training set, it seems that the proposed filter performs better when processing the signal after SAO. Thus, we choose LDSCA as the final embedding method and

TABLE 12. Performance comparison among different embedding methods. MSIM: multi-scale structural similarity (MS-SSIM).

Embedding methods	Y-PSNR	Y-MSIM
LCDSA	-2.54%	-2.31%
LDCSA	-2.81%	-2.62%
LDSCA	-2.78%	-2.61%
LDSAC	-2.67%	-2.78%
LCA	-2.24%	-2.00%

TABLE 13. BD-rate comparison between WCDANN-DSConv and WCDANN-Loop over VTM-11.0-NNVC anchor in RA configuration. MSIM: multi-scale structural similarity (MS-SSIM). Negative BD-rate values indicate coding gains.

Method	WCDANN-DSConv		WCDANN-Loop	
Metric	Y-PSNR	Y-MSIM	Y-PSNR	Y-MSIM
Class	BD-rate	BD-rate	BD-rate	BD-rate
C	-3.43%	-2.64%	-2.78%	-2.61%
D	-4.76%	-2.68%	-3.30%	-1.97%
Overall	-4.10%	-2.66%	-3.04%	-2.29%

TABLE 14. BD-rate comparison between WCDANN-DSConv and WCDANN-Loop over VTM-11.0-NNVC anchor in AI configuration. MSIM: multi-scale structural similarity (MS-SSIM). Negative BD-rate values indicate coding gains.

Method	WCDANN-DSConv		WCDANN-Loop	
Metric	Y-PSNR	Y-MSIM	Y-PSNR	Y-MSIM
Class	BD-rate	BD-rate	BD-rate	BD-rate
C	-4.58%	-3.99%	-4.62%	-4.56%
D	-4.97%	-3.70%	-4.69%	-4.03%
Overall	-4.78%	-3.85%	-4.66%	-4.30%

TABLE 15. BD-rate comparison between WCDANN-DSConv and WCDANN-Loop over VTM-11.0-NNVC anchor in LDP configuration. MSIM: multi-scale structural similarity (MS-SSIM). Negative BD-rate values indicate coding gains.

Method	WCDANN-DSConv		WCDANN-Loop	
Metric	Y-PSNR	Y-MSIM	Y-PSNR	Y-MSIM
Class	BD-rate	BD-rate	BD-rate	BD-rate
C	-3.95%	-3.12%	-2.81%	-2.68%
D	-4.72%	-3.17%	-3.05%	-1.98%
Overall	-4.34%	-3.15%	-2.93%	-2.33%

test it on C and D classes in AI, RA, and LDP configurations. The LDSCA is shown in Fig. 14 and test results are shown in Tables 13, 14 and 15. For clarity, we refer to WCDANN in the loop filter as WCDANN-Loop. Compared with VTM-11.0-NNVC, WCDANN-Loop achieves average 3.04%, 4.66% and 2.93% BD-rate reductions for Y channel on C and D classes, under RA, AI and LDP configurations, respectively.

It can be seen from the results that in AI configuration, WCDANN achieves similar performance to the post-processing filter in the loop filter, which shows that WCDANN performs better at I-frame. However, in RA and LDP configurations, compared to the post-processing filter, the performance of WCDANN is degraded when working in the loop filter. In RA and LDP configurations, the main types of compressed video frames are B-frames and P-frames, respectively. These frames are very different from I-frame and usually of low quality and have serious artifacts. Thus, more auxiliary information, such as partition and prediction, are required for these frames. Since the information of partition and prediction highly depends on compression artifacts and reconstruction distortion, it can be used to help the filter identify artifacts and distortions, thereby improving compressed image quality. Therefore, our further study includes using auxiliary information for WCDANN as input and extending WCDANN to the chroma channels by a guidance strategy.

V. CONCLUSION

We have proposed WCDANN for compression artifact removal that is a CNN post-processing filter based on depthwise separable convolution and attention mechanism. WCDANN consists of two novel modules WCDAB and RAB to effectively extract residual features from the input image. WCDANN adopts depthwise separable convolution in WCDAB and RAB to greatly reduce the amount of network parameters of the network. Moreover, WCDANN deploys two attention mechanisms of CAB and CSAB to emphasize important features of different modules. WCDANN is evaluated on CTC video sequences and compared with other existing methods to confirm its superiority in compression artifact removal. We have compared the performance of depthwise separable convolutions and standard convolutions, confirming the effectiveness of depthwise separable convolution. We have explored various methods to embed the proposed CNN filter inside the VVC loop and find the optimal embedding scheme. Experimental results show that WCDANN achieves average 2.81%, 4.12% and 3.81% BD-rate reductions over VTM-11.0-NNVC anchor for Y channel on A1, A2, B, C, D and E classes in RA, AI and LDP configurations, respectively.

REFERENCES

[1] B. Bross, J. Chen, S. Liu, and Y.-K. Wang, *Versatile Video Coding Editorial Refinements on Draft 10*, document JVET-T2001, Nov. 2020.

[2] *High Efficiency Video Coding*, document ISO/IEC 23008-2:2020, H.265, Information Technology, Aug. 2020.

[3] M. Karczewicz, N. Hu, J. Taquet, C.-Y. Chen, K. Misra, K. Andersson, P. Yin, T. Lu, E. François, and J. Chen, "VVC in-loop filters," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 10, pp. 3907–3925, Oct. 2021.

[4] B. Bross, Y.-K. Wang, Y. Ye, S. Liu, J. Chen, G. J. Sullivan, and J.-R. Ohm, ''Overview of the versatile video coding (VVC) standard and its applications,'' *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 10, pp. 3736–3764, Oct. 2021.

[5] O. Akbulut and M. Z. Konyar, ''Improved intra-subpartition coding mode for versatile video coding,'' *Signal, Image Video Process.*, vol. 16, no. 5, pp. 1363–1368, Jul. 2022.

[6] J. Xu, G. Wu, C. Zhu, Y. Huang, and L. Song, ''CNN-based fast CU partitioning algorithm for VVC intra coding,'' in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2022, pp. 2706–2710.

[7] C. Dong, Y. Deng, C. C. Loy, and X. Tang, ''Compression artifacts reduction by a deep convolutional network,'' in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 576–584.

[8] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, ''Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising,'' *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3142–3155, Jul. 2017.

[9] Y. Tai, J. Yang, X. Liu, and C. Xu, ''MemNet: A persistent memory network for image restoration,'' in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4539–4547.

[10] Z. Qi, C. Jung, and B. Xie, ''Subband adaptive image deblocking using wavelet based convolutional neural networks,'' *IEEE Access*, vol. 9, pp. 62593–62601, 2021.

[11] B. Xie, H. Zhang, and C. Jung, ''WCDGAN: Weakly connected dense generative adversarial network for artifact removal of highly compressed images,'' *IEEE Access*, vol. 10, pp. 1637–1649, 2022.

[12] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, ''SwinIR: Image restoration using Swin transformer,'' in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 1833–1844.

[13] R. Yang, M. Xu, T. Liu, Z. Wang, and Z. Guan, ''Enhancing quality for HEVC compressed videos,'' *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 7, pp. 2039–2054, Jul. 2019.

[14] W. Lin, X. He, X. Han, D. Liu, J. See, J. Zou, H. Xiong, and F. Wu, ''Partition-aware adaptive switching neural networks for post-processing in HEVC,'' *IEEE Trans. Multimedia*, vol. 22, no. 11, pp. 2749–2763, Nov. 2020.

[15] Z. Guan, Q. Xing, M. Xu, R. Yang, T. Liu, and Z. Wang, ''MFQE 2.0: A new approach for multi-frame quality enhancement on compressed video,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 3, pp. 949–963, Mar. 2021.

[16] J. Wang, M. Xu, X. Deng, L. Shen, and Y. Song, ''MW-GAN+ for perceptual quality enhancement on compressed video,'' *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 7, pp. 4224–4237, Jul. 2022.

[17] H. Huang, I. Schiopu, and A. Munteanu, ''Frame-wise CNN-based filtering for intra-frame quality enhancement of HEVC videos,'' *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 6, pp. 2100–2113, Jun. 2021.

[18] D. Ding, L. Kong, G. Chen, Z. Liu, and Y. Fang, ''A switchable deep learning approach for in-loop filtering in video coding,'' *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 7, pp. 1871–1887, Jul. 2020.

[19] S. Chen, Z. Chen, Y. Wang, and S. Liu, ''In-loop filter with dense residual convolutional neural network for VVC,'' in *Proc. IEEE Conf. Multimedia Inf. Process. Retr. (MIPR)*, Aug. 2020, pp. 149–152.

[20] Z. Huang, J. Sun, X. Guo, and M. Shang, ''One-for-all: An efficient variable convolution neural network for in-loop filter of VVC,'' *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 4, pp. 2342–2355, Apr. 2022.

[21] Y. Li, L. Zhang, and K. Zhang, ''Convolutional neural network based in-loop filter for VVC intra coding,'' in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2021, pp. 2104–2108.

[22] L. Wang, W. Jiang, X. Xu, and S. Liu, *AHG11: Neural Network Based in-Loop Filter*, document JVET-W0113, Jul. 2021.

[23] F. Zhang, C. Feng, and D. R. Bull, ''Enhancing VVC through CNN-based post-processing,'' in *Proc. IEEE Int. Conf. Multimedia Expo. (ICME)*, Jul. 2020, pp. 1–6.

[24] D. Ma, F. Zhang, and D. R. Bull, ''MFRNet: A new CNN architecture for post-processing and in-loop filtering,'' *IEEE J. Sel. Topics Signal Process.*, vol. 15, no. 2, pp. 378–387, Feb. 2021.

[25] T. Liu, W. Cui, C. Hui, F. Jiang, Y. Gao, and S. X. P. Wu, *AHG11: Post-Process Filter Based on Fusion of CNN and Transformer*, document JVET-Z0101, Apr. 2022.

[26] M. Santamaria, Y.-H. Lam, F. Cricri, J. Lainema, R. G. Youvalari, H. Zhang, M. M. Hannuksela, E. Rahtu, and M. Gaubbuj, ''Content-adaptive convolutional neural network post-processing filter,'' in *Proc. IEEE Int. Symp. Multimedia (ISM)*, Nov. 2021, pp. 99–106.

[27] M. Santamaria, J. Lainema, F. Cricri, R. G. Youvalari, H. Zhang, A. Zare, G. Rangu, H. R. Tavakoli, H. Afrabandpey, and M. Hannuksela, *AHG11: MPEG NNR Compressed Bias Update for the CNN Based Post-Filter of EE1-1.1*, document JVET-X0111, Oct. 2021.

[28] C. Bonnineau, W. Hamidouche, J.-F. Travers, N. Sidaty, and O. Deforges, ''Multitask learning for VVC quality enhancement and super-resolution,'' in *Proc. Picture Coding Symp. (PCS)*, Jun. 2021, pp. 1–5.

[29] F. Chollet, ''Xception: Deep learning with depthwise separable convolutions,'' 2016, *arXiv:1610.02357*.

[30] C. Dong, C. C. Loy, K. He, and X. Tang, ''Image super-resolution using deep convolutional networks,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2016.

[31] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, ''Swin transformer: Hierarchical vision transformer using shifted windows,'' in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.

[32] Y. Zhao, K. Lin, S. Wang, and S. Ma, ''Joint Luma and chroma multi-scale CNN in-loop filter for versatile video coding,'' in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, May 2022, pp. 3205–3209.

[33] B. Kathariya, Z. Li, H. Wang, and G. Van Der Auwera, ''Multi-stage locally and long-range correlated feature fusion for learned in-loop filter in VVC,'' in *Proc. IEEE Int. Conf. Vis. Commun. Image Process. (VCIP)*, Dec. 2022, pp. 1–5.

[34] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, ''Attention is all you need,'' 2017, *arXiv:1706.03762*.

[35] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, ''CBAM: Convolutional block attention module,'' in *Proc. Eur. Conf. Comput. Vis.*, Jul. 2018, pp. 3–19.

[36] E. Agustsson and R. Timofte, ''NTIRE 2017 challenge on single image super-resolution: Dataset and study,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 1122–1131.

[37] D. Ma, F. Zhang, and D. R. Bull, ''BVI-DVC: A training database for deep video compression,'' *IEEE Trans. Multimedia*, vol. 24, pp. 3847–3858, 2022.

[38] G. Bjøntegaard, *Calculation of Average PSNR Differences Between RD-Curves*, document VCEG-M33, 2001.

[39] *Video Coding Standardization GitLab*. Accessed: Jul. 7, 2023. [Online]. Available: https://vcgit.hhi.fraunhofer.de/jvet-ahg-nnvc/VVCSoftware_VTM/-/tree/VTM-11.0_nnvc

[40] S. Liu, A. Segall, E. Alshina, and R.-L. Liao, *JVET Common Test Conditions and Evaluation Procedures for Neural Network-Based Video Coding Technology*, document JVET-W2016, Apr. 2021.

[41] Z. Qi, C. Jung, Y. Liu, and M. Li, ''CNN-based post-processing filter for video compression with multi-scale feature representation,'' in *Proc. IEEE Int. Conf. Vis. Commun. Image Process. (VCIP)*, Dec. 2022, pp. 1–5.

[42] K. Cui, A. B. Koyuncu, A. Boev, E. Alshina, and E. Steinbach, ''Convolutional neural network-based post-filtering for compressed YUV420 images and video,'' in *Proc. Picture Coding Symp. (PCS)*, Jun. 2021, pp. 1–5.

HAO ZHANG received the B.S. degree in electronic information engineering from Changan University, China, in 2020. He is currently pursuing the M.S. degree in electronic engineering with Xidian University, China. His research interests include image fusion, video coding, and deep learning.

CHEOLKON JUNG (Member, IEEE) is a Born Again Christian. He received the B.S., M.S., and Ph.D. degrees in electronic engineering from Sungkyunkwan University, Republic of Korea, in 1995, 1997, and 2002, respectively. He was a Research Staff Member with the Samsung Advanced Institute of Technology, Samsung Electronics, Republic of Korea, from 2002 to 2007. He was also a Research Professor with the School of Information and Communication Engineering, Sungkyunkwan University, from 2007 to 2009. Since 2009, he has been with the School of Electronic Engineering, Xidian University, China, where he is currently a Full Professor and the Director of the Xidian Media Laboratory. His main research interests include image and video processing, computer vision, pattern recognition, machine learning, computational photography, video coding, virtual reality, information fusion, multimedia content analysis and management, and 3DTV.

DAN ZOU received the B.S. degree in communication engineering from Northwestern Polytechnical University, China, in 2017, and the M.S. degree in instrument science and engineering from Shanghai Jiao Tong University, China, in 2020. Since 2020, she has been a Standardization Engineer with Guangdong OPPO Mobile Telecommunications Corporation Ltd., China. Her research interests include video coding and multimedia communications.

MING LI received the B.S. degree in telecommunication engineering and the Ph.D. degree in communication and information systems from Xidian University, China, in 2005 and 2010, respectively. He was a Senior Research Staff in standardization with ZTE Corporation, China, from 2010 to 2019. Since 2019, he has been a Senior Standardization Engineer with Guangdong OPPO Mobile Telecommunications Corporation Ltd., China. His research interests include video coding and multimedia communications.