IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 17 May 2023, accepted 25 May 2023, date of publication 1 June 2023, date of current version 7 June 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3281975

RESEARCH ARTICLE

High Quality Video Frames From VVC: A Deep Neural Network Approach

TANNI DAS[1], KIHO CHOI[2,3], (Member, IEEE), AND JAEYOUNG CHOI[1], (Member, IEEE)

[1]School of Computing (AI-Software), Gachon University, Seongnam, Gyeonggi 13120, South Korea
[2]Department of Electronics Engineering, Kyung Hee University, Yongin, Gyeonggi 17104, South Korea
[3]Department of Electronics and Information Convergence Engineering, Kyung Hee University, Yongin, Gyeonggi 17104, South Korea

Corresponding authors: Kiho Choi (aikiho@khu.ac.kr) and Jaeyoung Choi (jychoi19@gachon.ac.kr)

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean Government [Ministry of Science and ICT (MSIT)] (NRF-2021R1F1A1060816) and was supported by the Gachon University research fund of 2021 (GCU-202110040001).

ABSTRACT In recent years, video content has become a significant contributor to Internet traffic, prompting the development of efficient codecs, such as High Efficiency Video Coding (HEVC) and Versatile Video Coding (VVC), to reduce bandwidth usage and storage requirements. However, these video coding standards still exhibit quality degradation and artifacts in the decoded frames. To address this issue, researchers have introduced several network architectures based on deep-learning algorithms; however, most of them focus on in-loop filtering, which requires additional bits to transmit filter information from the encoder to the decoder under a video-coding framework. In this paper, we propose a neural-network-based post-processing method to enhance the decoded frames. In the experimental result, the proposed model achieves a significant bitrate reduction, as measured by Bjøntegaard Delta of 4.54%, 4.13%, and 5.21% for random access (RA), low-delay (LD), and all-intra (AI) configurations, respectively, while also improving peak signal-to-noise ratio (PSNR).

INDEX TERMS VVC, post-processing, video compression, CNN.

I. INTRODUCTION

The use of video data has increased significantly in daily life, leading to an increase in the tension between the available transmission bandwidth and the vast amount of video content being consumed. Recent advancements in hardware technology have aimed to enhance the visual quality for users. For instance, high dynamic range (HDR), high frame rate (HFR), and ultrahigh definition (UHD) video formats with 4K and 8K resolutions have been introduced to provide a more realistic viewing experience [1]. Consequently, video codecs play a crucial role in reducing bitrates and producing compressed videos to alleviate traffic loads on transmission lines.

As video content continues to grow, efficient video codecs are required to ensure the high-quality display of compressed videos while delivering additional scene details found in new video formats, despite limited distribution networks. To this end, the joint video exploration team (JVET) developed by the ITU-T video coding expert group (VCEG) and the moving picture expert group (MPEG), established a new video coding standard called Versatile Video Coding (VVC) [2]. This new codec introduces fundamental compression techniques that demonstrate significant performance improvements over their predecessors. Specifically, VVC shows a 50% and 75% increase in compression efficiency for equal video quality compared with High Efficiency Video Coding (HEVC) [3] and Advanced Video Coding (AVC) [4], respectively.

Despite the improved performance of VVC compared with prior standards through the development of more effective tools and algorithms, reconstructed videos generated by VVC still suffer from several artifacts such as blockiness, blurriness, and ringing. The conventional block-based approach to video coding standards is primarily responsible for these artifacts. Additionally, the quantization of the transform coefficients contributes to blurriness and ringing artifacts, which worsen with increasing quantization parameter values. To address compression artifacts, recent studies have

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

employed handcrafted filters such as the deblocking filter (DBF), sample adaptive offset (SAO), and adaptive loop filter (ALF). Although the handcrafted nature of these filters shows improvement, there is still the potential to enhance the reconstruction quality.

The use of deep learning approaches, specifically deep convolutional neural networks (CNN), has significantly advanced the field of video compression and led to the development of CNN-based artifact-removal networks. These networks can be used as either in-loop filters (ILF) or for postprocessing (PP). Recent studies proposed the use of CNN-based in-loop filters [5], [6], [7], [8]. Whereas in-loop filters require consideration of both the encoding and decoding aspects, post-processing approaches are more flexible because they are applied after the decoding step. However, most CNN-based postprocessing architectures are designed for specific coding configurations, such as random access (RA), all-intra (AI), and low-delay (LD). Moreover, the justification for the trade-off between network complexity and performance remains inadequate.

In this paper, we present a novel CNN-based postprocessing framework for reducing the bit rate while maintaining the same reconstructed video quality. The proposed method addresses the problem of coding artifacts that arise in different quantization parameter (QP) scenarios by incorporating a QP map as prior information with the encoded input image, following the approach proposed in [8]. The main contributions of this study are summarized as follows:

1) We developed a single CNN-based post-processing network that can handle all RA, AI, and LD coding scenarios, thereby improving flexibility and reducing complexity.
2) We utilize minimal skip connections and a simple network architecture to further reduce the network complexity.
3) We optimized the QP map with the encoded input image to achieve better generalization of multi-QP-generated artifacts.

The remainder of this paper is organized as follows. Section II provides an overview of recent contributions to deep-learning-based postprocessing methods. The proposed method is described in Section III. Section IV presents the overall performance evaluation and analysis. Finally, Section V concludes the study.

II. RELATED WORKS

In the domain of video compression artifact removal, two approaches have been explored: ILFs that operate at both the encoder and decoder and out-of-loop post-processing algorithms designed at the decoder end. Several research endeavors have been conducted to overcome the limitations of traditional filters and achieve significant improvements over conventional video coding standards. In this section, several major studies on CNN-based filter architectures are reviewed. Because the approach proposed in this study primarily focuses on one of the aspects mentioned earlier (post-processing), these studies are categorized such that they reflect the overall research contribution flow in the direction of the proposed design inspiration in this study.

A. CNN-BASED IN-LOOP FILTER APPROACH

The research on the development of CNN-based ILF can be classified into three types. First, CNN models were devised as substitutes for the traditional filters. Second, CNN models are added after traditional filters. Finally, the CNN models are integrated with traditional filters. Park and Kim [9] proposed an in-loop filter CNN (IFCNN) that could replace SAO in HEVC. Their experiment showed BD rate reductions of 2.6% and 2.8% for the RA and LD configurations, respectively. Dai et al. proposed a variable filter size residual learning convolutional neural network (VRCNN) [10] that can replace both DB and SAO in HEVC. This variable filter size approach helps facilitate the variable block size transformation in HEVC, and residual learning is used to achieve faster convergence. The VRCNN has been reported to achieve an average BD rate reduction of 4.6%. Kang et al. proposed a multi-scale CNN (MMS-net) [11], which consists of two subnetworks with different scales that can replace DB and SAO in HEVC. This network is deeper and utilizes skip connections in each subnetwork with coding parameters to boost the restoration process. Wang et al. [8] proposed an attention-based dual-scale CNN (ADCNN) to replace conventional filters in a VVC. In this method, the quantization parameter and partitioning information are used as prior information and are adapted to different QPs. This method showed a gain for both the AI and RA configurations. A residual highway CNN was proposed in [12], in which the CNN network was included after the conventional filters in HEVC. This method consists of several residual units and convolution layers with a progressive training scheme for the QP bands. Wang et al. [13] suggested a neural network-based in-loop filter (CNNLF) consisting of two modules for image feature extraction and image quality improvement. The proposed in-loop filter was incorporated after the traditional filters in the VVC. Jia et al. [14] proposed a content-aware CNN that incorporates SAO and ALF. The experiment shows a 10.0% bitrate reduction in HEVC. Huang et al. [15] suggested a variable CNN (VCNN) that embeds an attention module into a residual block to extract informative features. This network can be added between DB and SAO for VVC.

B. CNN-BASED POST-PROCESSING APPROACH

For out-of-loop filters, the post-processing method is applied after the images are decoded. Several models have been proposed to increase the quality of the decoded images. Dong et al. [16] proposed an artifact removal CNN-based approach (AR-CNN) for JPEG-compressed images. In addition to a previously created super-resolution CNN (SRCNN) [17], an AR-CNN was implemented. According to previous reports, the ARCNN outperformed JPEG images by more than 1dB. A twenty layer CNN architecture with

residual learning was suggested by Li et al. [18]. To transfer the side information associated with video content complexity and quality indicators per frame, an up-to-one-byte flag was embedded in the bitstream to select a separate trained model in the post-processing module. Zhang et al. [19] proposed a CNN-based post-processing architecture for VVC compressed video sequences that utilized 16 identical residual blocks and occupied three types of skip connections. As an extension of [19], in [20], a generative adversarial network (GAN)-based training strategy was applied to improve the quality of VVC-compressed reconstructed images. Two training methodologies were applied to obtain a significant improvement in perceptual quality. The generator was first trained using the mean absolute difference loss, and both the generator and discriminator were jointly trained based on perceptually inspired loss functions. Bonnineau et al. [1] proposed a multitask learning-based method that influences the similarity of super-resolution and quality enhancement tasks by sharing parameters with a single shared network and task-specific modules. As mentioned previously, the QP map was concatenated with the encoded image to improve the generalization of the model with different quantization parameters.Wang et al. [21] proposed a CNN-based single model employing QP and partition information to improve multiquality reconstruction and quality enhancement, respectively. A three-branch network was proposed to process the three different components. Meng et al. [22] proposed a quality enhancement network for VVC-compressed videos. This network consists of a fusion subnet and an enhancement subnet that exploit the temporal motion and spatial detail, respectively. In [23], an image restoration network was proposed in which multi-scale spatial priors were used to extract multi-scale features. Four residual blocks were applied to obtain the high-dimensional features.

C. PERSPECTIVE FROM PRIOR WORKS

In our research, our objective was to improve the quality of the decoded frame while simultaneously reducing the bitrate. While most previous studies have developed in-loop filtering networks for the HEVC codec, we propose a network that can be implemented on the decoder side after frame reconstruction. This post-filtering approach provides several advantages over in-loop filtering, including the following:

1) **Reduced computational complexity**: Compared to the in-loop filter, the post-processing filter offers a reduced computational burden as it is applied only once during the decoding process. This significantly improves the efficiency of real-time video applications by lowering the computational requirements and memory usage.
2) **Enhanced visual quality**: By applying the post-processing filter, the visual quality of the video can be noticeably improved. It effectively reduces noise, enhances edge sharpness, and improves contrast. Additionally, it helps rectify any encoding errors, resulting in a more faithful representation of the original scene.
3) **Bitrate reduction**: The post-processing filter contributes to bitrate reduction by eliminating unnecessary details and smoothing out noise in the video. This leads to a more efficient utilization of bandwidth and storage resources without compromising visual quality.
4) **Implementation simplicity**: Implementing the post-processing filter is typically simpler compared to the in-loop filter. With only a single pass required during the decoding process, it reduces development complexity and expedites the time-to-market for video applications. This advantage allows for faster deployment and easier integration into existing video coding frameworks.

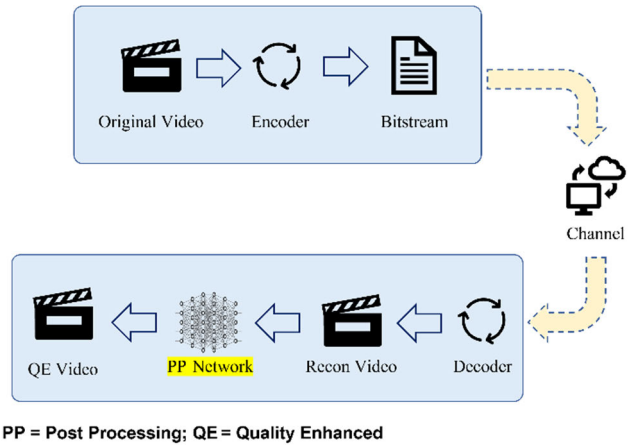

FIGURE 1. Integrated deep learning based post-processing approach in typical coding workflow. Yellow color indicates the focus of this research.

It is worth noting that the majority of previous researchers have focused on developing their networks for in-loop filtering, which is integrated within the HEVC codec. In contrast, our network is specifically designed to be applied at the decoder side, enhancing the frame after the reconstruction process. Furthermore, we employed the latest VVC codec for all three coding configurations (i.e., RA, LD, and AI).

III. PROPOSED METHOD

In this section, the proposed algorithm and architecture are presented and explained. Figure 1 illustrates a CNN-based post-processing approach integrated into a conventional processing pipeline. In this post-processing pipeline, the transmitted bitstream is decoded to produce reconstructed frames, and a CNN filter is applied to enhance the video quality of the reconstructed frames. Based on the conventional processing pipeline, the proposed algorithm investigates a method for utilizing QP information and improving the feature flow through a deep network. Specifically, the proposed method comprises functions that use QP-based prior coding information (i.e., QP map) and a network architecture based on deep learning technology. Detailed information is provided in this section.

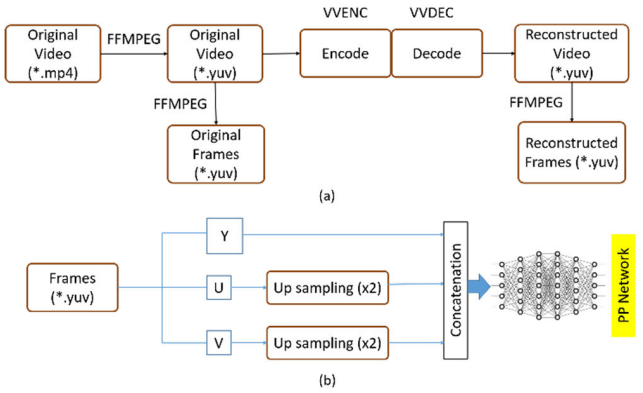

FIGURE 2. Overview of the proposed framework. (a) Dataset preparation in YUV format, and (b) Conversion between YUV 4:2:0 format and YUV 4:4:4 format before network input.

A. FRAMEWORK

The initial stage of the proposed framework involves dataset preparation, as shown in Figure 2(a). To construct the training dataset, commonly used videos from [24] in the MP4 format were used. To facilitate the integration of the proposed method with raw-format videos, MP4 videos were converted into YUV format using FFmpeg [25]. Following the conversion process, the original videos in YUV format were subjected to VVenC [26] and VVdeC [27] to produce reconstructed videos in YUV format. Both the original and processed images were reconstructed using FFmpeg. To process videos in a 4:2:0 format, the chroma components (i.e., U and V channels) of both the original and reconstructed frames were upscaled and fed into the proposed postprocessing network. Figure 2(b) illustrates the procedure for inputting data into the network.

B. QP MAP

Feeding the QP map to the networks is a crucial component of the proposed method and contributes significantly to the generation of high-quality outputs. QP determines the quantization step, which in turn influences the quality of the reconstructed video frames.

An increase in the quantization parameter leads to a higher distortion because a coarser quantization step is applied to transform the coefficients with a larger QP. This results in the loss of most of the high-frequency information and a wider distribution range for the compensation value between the reconstructed and original pixels. To enhance the network's ability to compensate for this distribution range, prior information in the form of a QP map is necessary to produce reconstructed outputs that are as close as possible to the input. To significantly filter inputs with varying qualities and improve the model's ability to generalize across multiple quantization parameters, we integrated the QP map as prior information into the network. When we concatenate QP-map as prior information with the reconstructed frame with varying quality, it helps introducing the distortion diversity associated with each individual QP value to the network. The network utilizes this prior knowledge to get an idea about the distortion level associated with each QP value. Thus, QP-map increases the network's ability to generalize across multiple QP values and helps to get the network output as close as possible to the original frame. The QP map is fed into a network of the same size as that of the reconstructed input frame.

The QP map generates a normalized value computed as in (1).

$$QP_{map}(u, v) = \frac{QP(u, v)}{QP_{max}} \tag{1}$$

where, $u = 1, \ldots W$ and $v = 1, \ldots H$ denote the horizontal and vertical pixel coordinates, respectively. For the VVC, the QP_{max} value was 63, where QP_{max} specified the maximum amount of compression that could be applied to each coding unit in a frame.

C. NETWORK ARCHITECTURE

Recently, NN-based architectures have played an important role in enhancing the quality of the reconstructed frames from various video coding standards. Researchers have experimented with various network architectures to further improve frame quality. However, video frames are more complex than still images in terms of motion and temporal dependency, making it challenging to achieve the same quality as that of the original frames after reconstruction. To address this issue, complex network architectures have been proposed. However, integrating these architectures with conventional video codecs and implementing them in real-world scenarios remains challenging. Based on earlier studies, this study aims to develop a lightweight CNN model that satisfies three key criteria: a) easy to implement and lightweight, b) generalized across different QP values with varying input information, and c) generalized for different coding configurations, such as RA, LD, and AI.

The proposed network architecture for enhancing the quality of the reconstructed frames is shown in Figure 3. The input comprises a VVC-decoded frame and a QP map. The proposed network comprises three main parts: a) Forward block, b) Feature extraction block, and c) Tail block. Furthermore, the proposed architecture leveraged the benefits of residual connections. The first part of the proposed network, that is the forward block, receives the concatenation of these inputs.

The proposed network utilizes a QP map with identical dimensions to the reconstructed frame, which is concatenated with the frame and fed into a forward block. This allows the QP map to play the role as a form of prior coding information for the network, enabling it to produce higher-quality filters despite variations in frame quality.

The proposed forward block of the network comprises a projection and activation layer. The projection layer was designed as a 1×1 convolution filter with 128 channels, which reduced the dimensionality and number of feature maps while retaining essential features. Each channel can

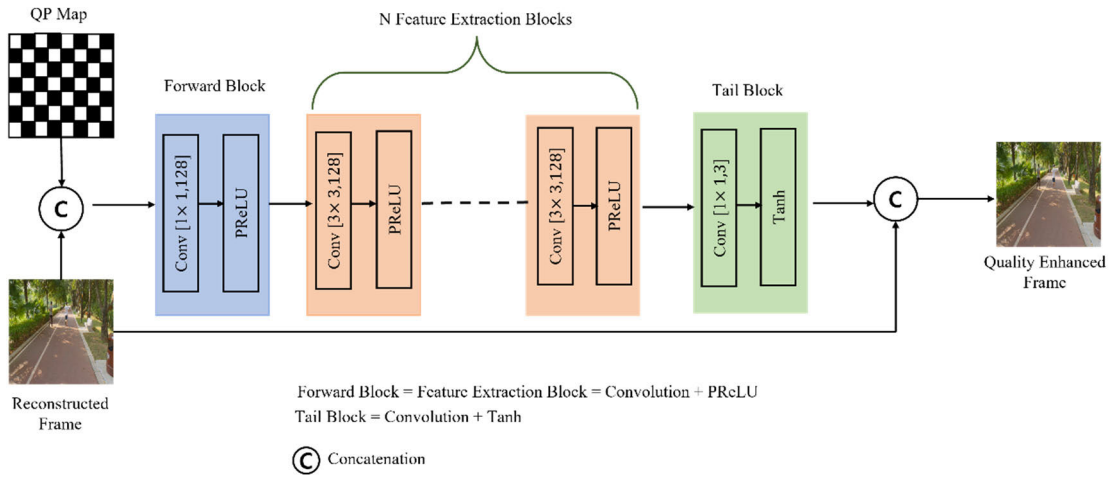

FIGURE 3. Proposed CNN architecture for post-processing.

capture distinct characteristics from the input, thus enabling the network to learn global features. The inclusion of 128 channels allowed the learning of multiple feature representations. A parametric rectified linear unit (PReLU) was used for the activation layer in this block. The feature maps generated by the forward block are subsequently forwarded to the feature-extraction block.

Multiple feature extraction blocks can be utilized in the proposed network by considering the tradeoff between model complexity and performance. To extract features more effectively while minimizing the network parameters and complexity, we employed 16 feature extraction blocks in our experiments. These blocks followed the same structure as the forward block, except for the filter size. A 3×3 kernel was used in the feature extraction blocks to capture more contextual information from the previous layers. Following the convolution layer, the PReLU activation function was applied. Because of the variation in nonlinearity across different layers, with deeper layers being more nonlinear than earlier layers, PReLU was employed to facilitate convergence in the deeper layers of the network.

In the final stage of the network, which is the tail block, the architecture is similar to that of the forward block, except for the activation function. The hyperbolic tangent (Tanh) was used as the activation function to achieve a more accurate mapping of the tail block output. Because the proposed network learns in a residual form, a skip connection between the input and output is included. This skip connection ensures an uninterrupted gradient flow and smooth propagation of information from earlier layers.

The mapping between the input (i.e., reconstructed frame and QP map) and output (i.e., original frame) is expressed as follows:

$$y = H_\theta(\hat{F} \oplus QP_{map}) \oplus \hat{F} \tag{2}$$

where y is the output or original frame, H_θ is the operation of CNN architecture; $\hat{F}$ is the input or decoded frame; $\oplus$ expresses the concatenation operation.

The network architecture presented in this study is unique in that sense it incorporates a 1×1 convolution filter at the outset, followed by a block of 3×3 convolution filters, and ultimately another 1×1 convolution filter. The rationale for this approach was based on careful consideration of several factors.

1) **Non-linearity**: Incorporating a 1×1 convolution filter alongside the 3×3 convolution filter introduces non-linearity into the model, enabling it to capture complex features and intricate relationships within the input data. This integration of different filter sizes enhances the model capacity to learn and represent more intricate patterns, resulting in improved performance.
2) **Computation reduction**: Leveraging the 1×1 convolution filter allows for a reduction in the number of input channels while preserving the spatial dimensions of the input tensor. By decreasing the dimensionality of the input, the subsequent 3×3 convolution filter becomes computationally more efficient, enabling faster processing and reduced computational cost without sacrificing valuable information.
3) **Expanded receptive field**: Compared to a single 1×1 convolution filter, the 3×3 convolution filter has a larger receptive field, enabling it to capture and integrate more extensive spatial information. This broader scope facilitates the ability to capture global features and long-range dependencies, enhancing its performance in handling complex visual tasks that require a comprehensive understanding of the input data.
4) **Enhanced accuracy**: The combined utilization of the 1×1 and 3×3 convolution filters enables the model to learn and incorporate both low-level and high-level features. This multi-scale feature extraction enhances the accuracy and effectiveness in various computer vision tasks, particularly those related to visual quality improvement. By leveraging the strengths of both

TABLE 1. Details of the BVI-DVC dataset.

Video Resolution	Sequence Number	Bit Depth	Chroma Sampling
480x272	200	10	4:2:0
960x544	200	10	4:2:0
1920x1088	200	10	4:2:0
3840x2176	200	10	4:2:0

filters, our proposed architecture achieves improved performance and superior accuracy compared to previous methods.

D. TRAINING CONFIGURATION

The BVI-DVC [24] dataset was used to train the model. This dataset comprised 800 videos of varying resolutions ranging from 270p to 2160p, providing a diverse set of training data. The videos were compressed under the JVET neural network-based video coding (NNVC) common test condition (CTC) [28] using the RA, LD, and AI configurations. The videos were in MP4 format and converted into YUV format with chroma sampling 4:2:0 and a bit depth of 10, as described in Subsection III-A. To simplify the process, 10 frames were extracted from each video, resulting in a training dataset of 8,000 frames. The chroma channels were up-sampled by a factor of two to match the spatial resolution of the Luma channel, because the proposed network cannot handle different input sizes. A random patch of size 240 $\times$ 240 from each frame was selected as the input, and horizontal and vertical flips were applied as data augmentation techniques. Table 1 presents a detailed description of the training dataset.

We trained five distinct models with the same network architecture based on different QP values. The QP values used were 22, 27, 32, 37, and 42 in accordance with the JVET NNVC CTC guidelines. These models were then used in the subsequent evaluation stage for different base QP values, and the same model-generation strategy was used for the RA, LD, and AI configurations. Each CNN model was trained for 200 epochs using the Adam optimizer with a learning rate of 10^{-4}, and hyper-parameters of $\beta_1 = 0.9$ and $\beta_2 = 0.999$ were utilized for calculating averages of the gradient during the learning process.

$$CNN\ Models = \begin{cases} Model_{Qp=22}, & QP_{base} < 24.5 \\ Model_{Qp=27}, & 24.5 \leq QP_{base} < 29.5 \\ Model_{Qp=32}, & 29.5 \leq QP_{base} < 34.5 \\ Model_{Qp=37}, & 34.5 \leq QP_{base} < 39.5 \\ Model_{Qp=42}, & QP_{base} \geq 39.5 \end{cases} \tag{3}$$

The L_2 loss function is used as the loss function, which is given in (4).

$$L_2 orMSE = \sum_{i=1}^{n} \left(y_i - \hat{F}_i\right)^2 \tag{4}$$

where y_i and $\hat{F}_i$ respectively represent the original and enhanced pixel values after applying the CNN filter.

TABLE 2. BD-rate for random access configuration.

Class	Sequence	BD-BR (%)		
		Y	U	V
A1	Tango	-3.36%	-11.50%	-14.91%
	FoodMarket	-2.93%	-8.33%	-9.80%
	Campfire	-3.89%	-8.20%	-18.02%
A2	CatRobot	-4.98%	-18.36%	-18.33%
	DaylightRoad	-5.61%	-14.32%	-14.49%
	ParkRunning	-1.47%	-7.73%	-6.97%
B	MarketPlace	-2.45%	-13.76%	-13.36%
	RitualDance	-4.33%	-11.99%	-14.30%
	Cactus	-4.38%	-12.59%	-11.70%
	BasketballDrive	-5.35%	-15.51%	-18.88%
	BQTerrace	-7.05%	-13.07%	-12.79%
C	BasketballDrill	-5.99%	-12.12%	-18.35%
	BQMall	-6.47%	-16.57%	-19.18%
	PartyScene	-5.75%	-15.18%	-13.34%
	RaceHorses	-4.02%	-15.78%	-18.51%
	Overall	**-4.54%**	**-13.00%**	**-14.86%**

IV. EXPERIMENTAL RESULTS AND DISCUSSION

A. EVALUATION PROCESS

To evaluate the effectiveness of the proposed method, the JVET NNVC CTC sequences were selected for evaluation and were not included in the training dataset. These sequences were excluded from the training dataset and comprised 18 sequences categorized into classes A1, A2, B, C, and E. The RA, LD, and AI configurations were evaluated using Class A1, A2, B, and C; Class B, C, and E; and Class A1, A2, B, C, and E, respectively, as prescribed in the JVET NNVC CTC. The test QP values for all the configurations were 22, 27, 32, 37, and 42.

B. EVALUATION METRIC

The proposed method was assessed by comparing the quality of the output generated by the proposed network with that of the decoded frames produced by the VVenC software using the peak signal-to-noise ratio (PSNR) as the evaluation metric. The equation for the PSNR is presented in (5).

$$PSNR\_Y = 10 * log10\left(\frac{(255 \ll (bitDepth - 8))^2}{MSE}\right) \tag{5}$$

where *bitDepth* describes the number of bits used to represent each pixel in an image as well as the color information that is stored there. MSE stands for mean squared error that assesses the average squared difference between the observed and predicted values. The $\ll$ represents a left-shift operator which is used to calculate the maximum value at the given *bitDepth*.

C. EXPERIMENTAL SETUP

The test was conducted using PyTorch [29] as the deep-learning framework on an Ubuntu operating system. The hardware configuration comprised two AMD EPYC 7513 32-Core CPUs, 384 GB of RAM, and an NVIDIA A6000 GPU. The training process for each QP value with 200 epochs required approximately 28 h.

TABLE 3. BD-rate for low delay configuration.

Class	Sequence	BD-BR (%)		
		Y	U	V
B	MarketPlace	-2.17%	-21.17%	-16.79%
	RitualDance	-2.77%	-14.39%	-15.43%
	Cactus	-3.28%	-18.48%	-17.55%
	BasketballDrive	-3.30%	-17.53%	-19.51%
	BQTerrace	-4.00%	-19.93%	-18.06%
C	BasketballDrill	-4.15%	-18.88%	-23.35%
	BQMall	-6.30%	-22.11%	-23.11%
	PartyScene	-4.60%	-25.53%	-20.50%
	RaceHorses	-3.17%	-21.44%	-26.91%
E	FourPeople	-5.97%	-11.26%	-15.07%
	Johnny	-5.57%	-11.45%	-21.02%
	KristenAndSara	-4.25%	-11.60%	-16.56%
	Overall	**-4.13%**	**-19.94%**	**-20.13%**

TABLE 4. BD-rate for all intra configuration.

Class	Sequence	BD-BR (%)		
		Y	U	V
A1	Tango	-4.15%	-11.63%	-13.75%
	FoodMarket	-4.12%	-8.39%	-8.56%
	Campfire	-2.45%	-3.39%	-12.20%
A2	CatRobot	-7.61%	-18.26%	-16.97%
	DaylightRoad	-8.21%	-15.71%	-11.73%
	ParkRunning	-1.52%	-1.73%	-1.98%
B	MarketPlace	-3.60%	-12.25%	-13.66%
	RitualDance	-5.80%	-13.02%	-14.88%
	Cactus	-4.75%	-7.00%	-8.91%
	BasketballDrive	-4.09%	10.51%	-2.46%
	BQTerrace	-4.51%	6.13%	8.96%
C	BasketballDrill	-6.71%	-4.69%	-15.27%
	BQMall	-6.74%	-6.52%	-11.30%
	PartyScene	-4.02%	3.44%	3.36%
	RaceHorses	-3.74%	-11.37%	-18.38%
E	FourPeople	-7.86%	-7.42%	-10.80%
	Johnny	-7.39%	-8.19%	-15.28%
	KristenAndSara	-16.52%	-7.44%	-8.67%
	Overall	**-5.21%**	**-6.50%**	**-9.58%**

D. COMPRESSION PERFORMANCE ANALYSIS

Tables 2, 3, and 4 summarize the compression performance of the proposed architecture for the RA, LD, and AI configurations. The Bjontegaard Delta bit rate (BD-BR) [30] metric is employed by JVET to assess the reduction in bit rate. This metric proves valuable for comparing the coding efficiency of various video codecs or encoding settings since it considers both the bitrate and video quality. By quantifying the disparity in bitrate required to achieve an equivalent quality level between two different codecs, the BD-BR metric enables an objective evaluation of compression efficiency. A lower BD-BR value indicates a superior coding efficiency. The results in the tables indicate that the proposed method consistently achieved significant coding gains for all test sequences. Specifically, Table 2 shows that the proposed method achieves overall coding gains of 4.54%, 13.00%, and 14.86% for the Luma (i.e., Y component) and Chroma (i.e., U and V components), respectively, compared with the VVC compressed contents. Notably, Class B and Class C sequences showed significant BD rate savings, particularly at lower resolutions. For example, the BQTerrace sequence with a resolution of 1920 × 1080 and the BQMall sequence with a resolution of 832 × 480 exhibited BD-rate savings of 7.05% and 6.47%, respectively, for Luma.

Table 3 lists the coding performances of the proposed architecture for the LD configuration. The results showed overall BD-rate reductions of 4.13%, 19.94%, and 20.13% for the Luma and Chroma components, respectively. The LD configuration performs well in Class C sequences with lower resolutions, while also showing improved results in high-resolution sequences. Specifically, the BQMall sequence in Class C exhibited a coding gain of 6.30% for Luma, demonstrating the adaptability of the proposed network to lower-resolution sequences.

Table 4 lists the coding performances of the proposed architecture for the AI configuration. The results showed improved coding gains of 5.21%, 6.50%, and 9.58% for Luma and Chroma components, respectively, compared to the VVC compressed content. This significant performance improvement highlights the effectiveness of the proposed method. Moreover, the proposed method achieves considerable BD-rate reduction for both high-resolution and low-resolution image sequences.

TABLE 5. Comparison with state-of-the-art method for RA configuration.

Methods	Channels		
	Y	U	V
Zhang et. al. [20]	-3.43%	N/A	N/A
Ours	-4.54%	-13.00%	-14.86%

To evaluate the effectiveness of the proposed method, we compared it with the latest research introduced in [20], and the results are presented in Table 5. Specifically, the comparison was performed in the RA configuration, aligning with the CTC of JVET. In terms of coding efficiency measured by BD-BR, the proposed method outperformed [20] with a value of −4.54% using a less complex deep learning model architecture, while [20] reported a BD-BR value of −3.43% for the Y channel. Notably, the strength of the proposed method was evident in the results for class D, which represents the most challenging class due to its low resolution (i.e., 416 × 240). In this case, our proposed model surpassed [20], achieving a Y channel bit reduction of 6.64% compared to the reported rate of 5.80%. Furthermore, it is worth mentioning that our proposed method comprehensively addressed all three channels (i.e., Y, U, V) using a single network, while the network presented in [20] could only handle the Y component.

E. VISUAL QUALITY EVALUATION

Figure 4 to 6 show the comparative visual quality for the RA, LD, and AI configurations between the VVC-compressed content and the network results, with the visual quality

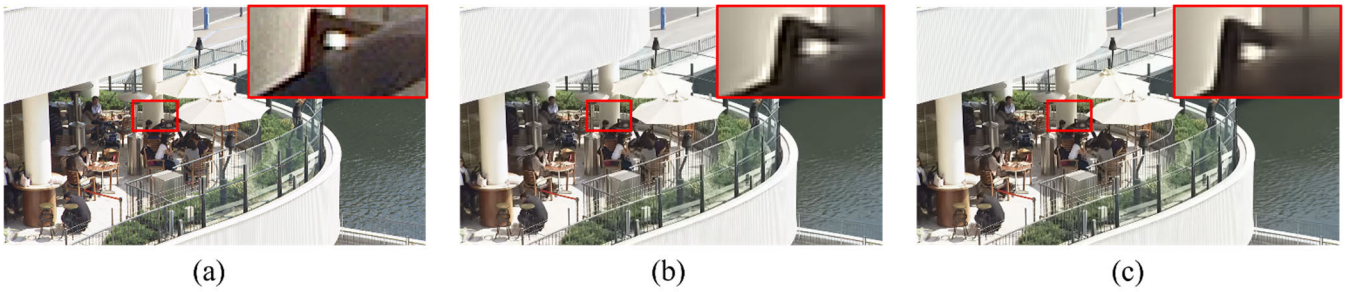

FIGURE 4. Example sequence of BQTerrace (a) original, (b) VVC compressed and (c) proposed approach for RA configuration with QP 42.

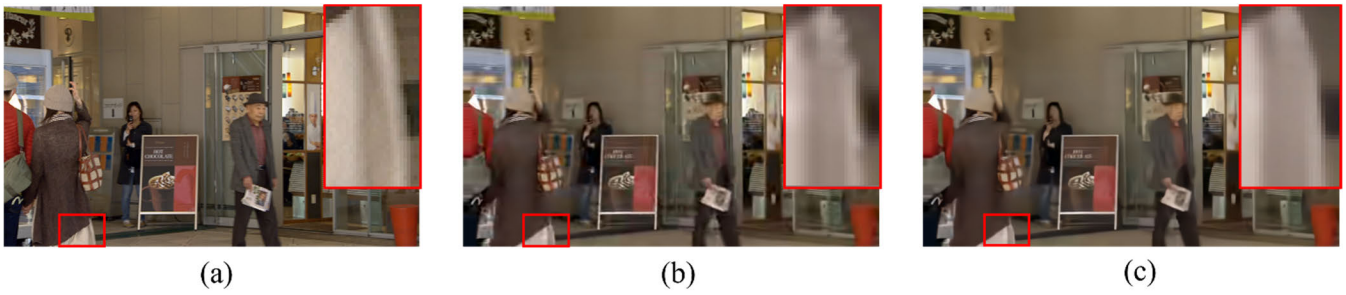

FIGURE 5. Example sequence of BQMall (a) original, (b) VVC compressed and (c) proposed approach for LD configuration with QP 42.

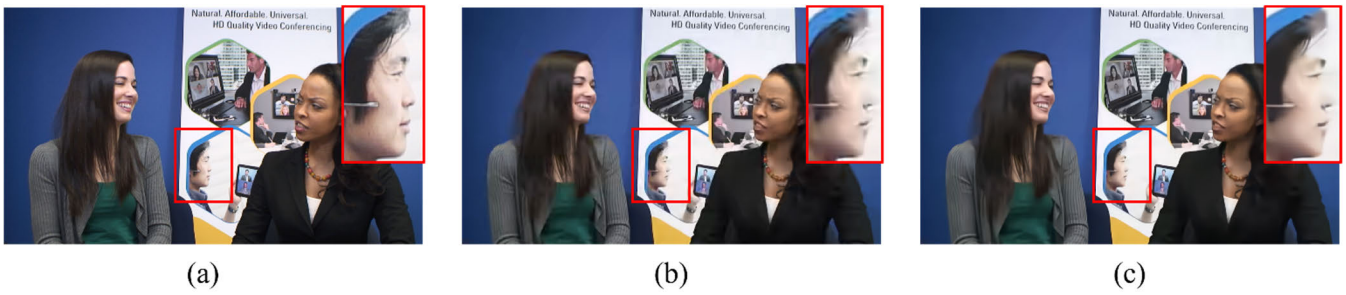

FIGURE 6. Example sequence of KristenAndSara (a) original, (b) VVC compressed and (c) proposed approach for AI configuration with QP 42.

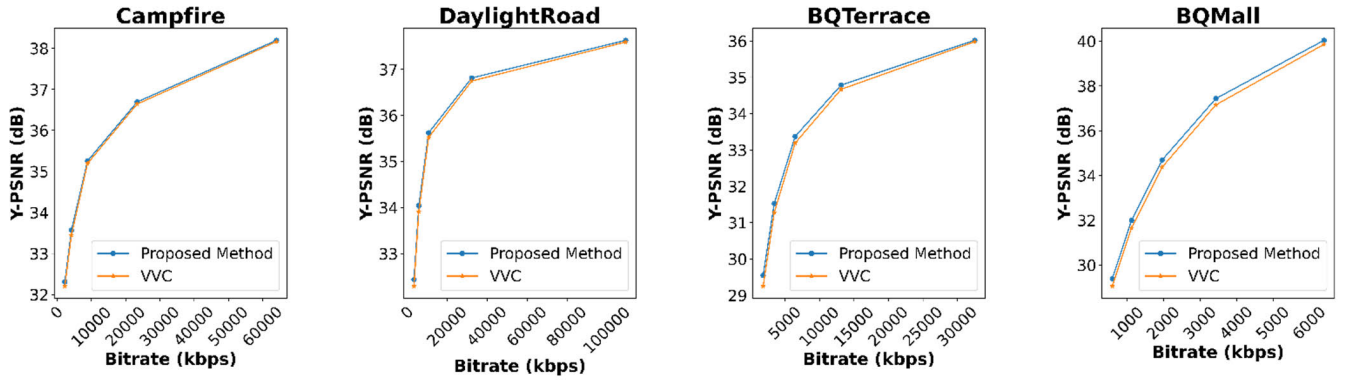

FIGURE 7. PSNR curves of selected sequences for RA configuration. Orange curve stands for VVC while blue curve stands for the proposed method.

assessment being conducted at a high QP value of 42. The proposed network output in Figure 4 displays smoother edges for the BQTerrace sequence with a resolution of 1920 × 1080, exhibiting a 0.29 dB PSNR gain over VVC. In Figure 5, the LD configuration shows more textural detail than the codec, particularly in the white skirt of the BQMall sequence, with a 0.35 dB PSNR gain. Figure 6 shows that the AI configuration achieves a 0.7 dB gain and displays less noticeable blocking artifacts than the VVC. Notably, the results for the RA, LD, and AI configurations were observed at the highest compressed parameter of QP 42, which presents challenging coding compression. Despite this, the proposed network

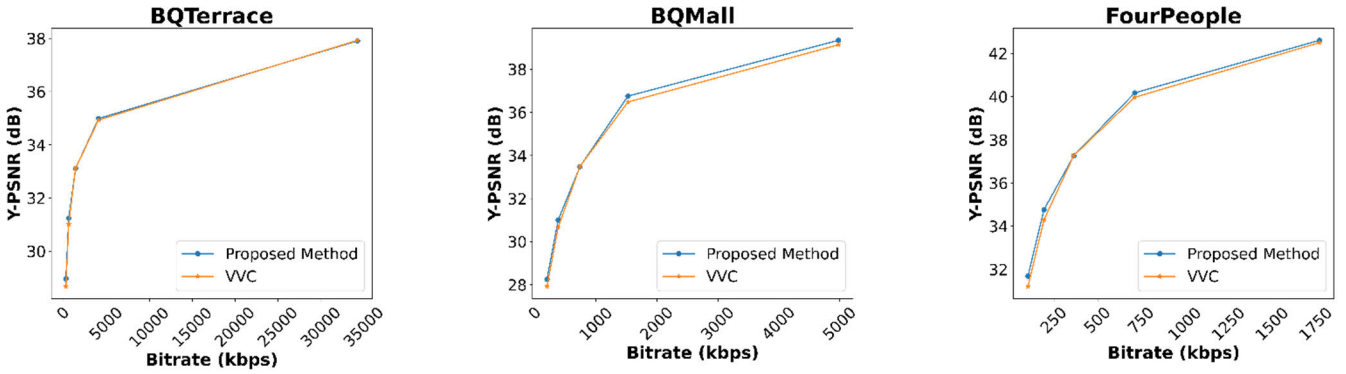

FIGURE 8. PSNR curves of selected sequences for LD configuration. Orange curve stands for VVC while blue curve stands for the proposed method.

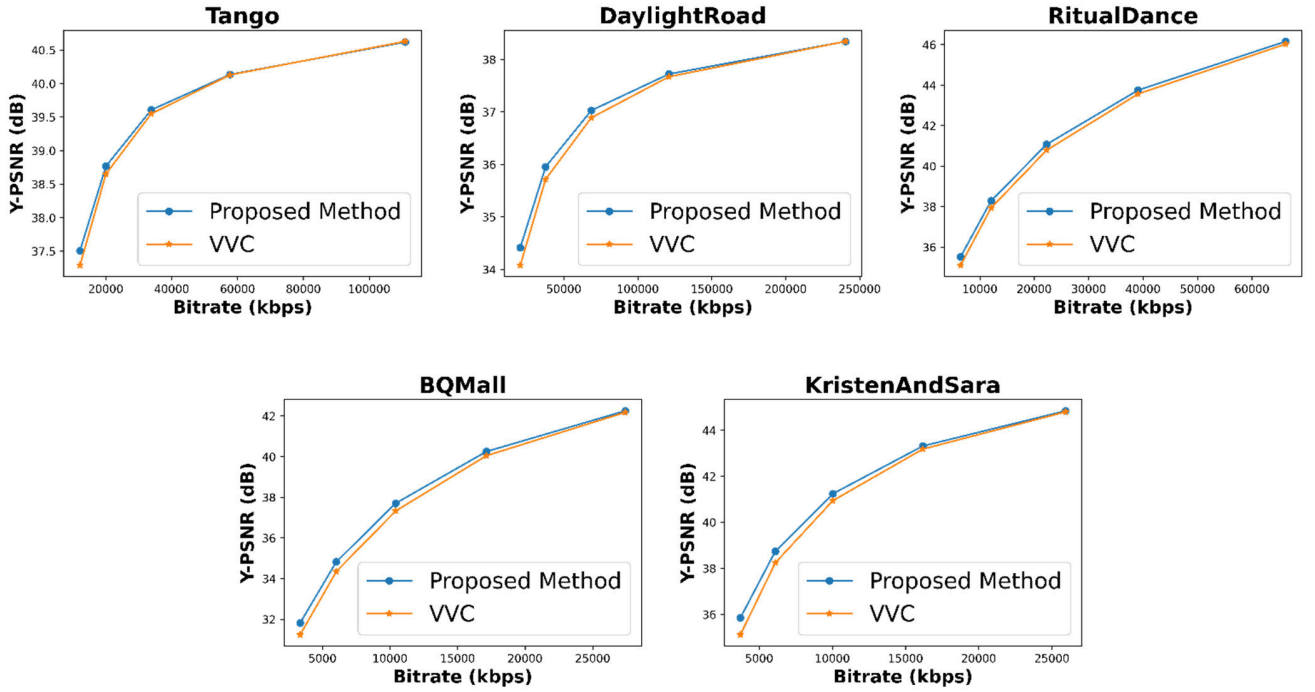

FIGURE 9. PSNR curves of selected sequences for AI configuration. Orange curve stands for VVC while blue curve stands for the proposed method.

significantly reduces coding artifacts and enhances quality while reducing the bitrate.

F. RATE-DISTORTION PLOT ANALYSIS

To assess RD performance, we present video sequences from the RA, LD, and AI configurations in Figure 7 to 9. Figure 7 shows four sequences with varying resolutions for the RA scenario, indicating that the proposed network performed well for lower-resolution sequences. Similarly, Figures 8 and 9 show three and five sequences from the JVET CTC for the LD and AI scenarios, respectively. In both cases, the proposed method showed improved results across the five QP levels and different-resolution video sequences. When examining Figure 7 to 9, it becomes evident that the rate-distortion (RD) curves of the proposed method and the original codec overlap at higher bitrates. This occurrence can be attributed to the increased available encoding space for the video, resulting in reduced compression and, subsequently, diminished distortion. Consequently, at high bitrates, the RD curve of the proposed method aligns with that of the original encoder. This phenomenon arises because the proposed method excels in enhancing the decoded frames at low to moderate bitrates, where compression artifacts are more pronounced. However, at high bitrates, where the original codec already produces high-quality videos, the need for enhancement diminishes, leading to the overlapping of RD curves. Notably, our network demonstrates its strength in further improving the reconstructed frame, particularly for higher QP values.

G. DISCUSSION

The results of the proposed post-processing filter approach demonstrated a significant reduction in artifacts associated

with the latest video coding standard, VVC, across various QP values ranging from 22 to 42. Specifically, the proposed method exhibited coding gains for the RA, LD, and AI scenarios. Moreover, the single postfilter architecture utilized in our approach maintained a simple and hardware-friendly design, which led to faster inference times. Although we used the YUV format as the input in our proposed method, there was a limitation in the conversion process during the preparation of the training data. In future work, we intend to enhance the ability of the network to handle the YUV format more efficiently by considering the formatting conversion process during the network design, while maintaining its simplicity.

V. CONCLUSION

In this study, we propose a novel CNN-based post-processing filter approach for reconstructed videos. The proposed method utilizes a QP map to generate inputs with varying frame qualities and derives the optimal number of feature extraction blocks with minimal skip connections for faster inference on low-end hardware. To demonstrate the effectiveness of the proposed approach, a single network was tested on three different video configurations (i.e., RA, LD, and AI) with five QP values for each configuration. The experimental results show that the proposed single-network architecture outperforms VVC in terms of BD-rate reduction for all three configurations with five different QP values, while maintaining a simple network architecture.

REFERENCES

[1] C. Bonnineau, W. Hamidouche, J.-F. Travers, N. Sidaty, and O. Deforges, "Multitask learning for VVC quality enhancement and super-resolution," May 2021, *arXiv:2104.08319*.

[2] 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, doi: 10.1109/TCSVT.2021.3101953.

[3] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1649–1668, Dec. 2012, doi: 10.1109/TCSVT.2012.2221191.

[4] T. Wiegand, G. J. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 7, pp. 560–576, Jul. 2003, doi: 10.1109/TCSVT.2003.815165.

[5] S. Bouaafia, S. Messaoud, R. Khemiri, and F. E. Sayadi, "VVC in-loop filtering based on deep convolutional neural network," *Comput. Intell. Neurosci.*, vol. 2021, Jul. 2021, Art. no. e9912839, doi: 10.1155/2021/9912839.

[6] C. D. K. Pham, C. Fu, and J. Zhou, "Deep learning based spatial–temporal in-loop filtering for versatile video coding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 1861–1865, doi: 10.1109/CVPRW53098.2021.00206.

[7] Z. Huang, Y. Li, and J. Sun, "Multi-gradient convolutional neural network based in-loop filter for VVC," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2020, pp. 1–6, doi: 10.1109/ICME46284.2020.9102826.

[8] M.-Z. Wang, S. Wan, H. Gong, and M.-Y. Ma, "Attention-based dual-scale CNN in-loop filter for versatile video coding," *IEEE Access*, vol. 7, pp. 145214–145226, 2019, doi: 10.1109/ACCESS.2019.2944473.

[9] W.-S. Park and M. Kim, "CNN-based in-loop filtering for coding efficiency improvement," in *Proc. IEEE 12th Image, Video, Multidimensional Signal Process. Workshop (IVMSP)*, Jul. 2016, pp. 1–5, doi: 10.1109/IVMSPW.2016.7528223.

[10] Y. Dai, D. Liu, and F. Wu, "A convolutional neural network approach for post-processing in HEVC intra coding," in *Proc. Int. Conf. Multimedia Modeling*, Jan. 2017, pp. 28–39, doi: 10.1007/978-3-319-51811-4_3.

[11] J. Kang, S. Kim, and K. M. Lee, "Multi-modal/multi-scale convolutional neural network based in-loop filter design for next generation video codec," in *Proc. Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 26–30, doi: 10.1109/ICIP.2017.8296236.

[12] Y. Zhang, T. Shen, X. Ji, Y. Zhang, R. Xiong, and Q. Dai, "Residual highway convolutional neural networks for in-loop filtering in HEVC," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3827–3841, Aug. 2018, doi: 10.1109/TIP.2018.2815841.

[13] Y. Wang, J. Zhang, Z. Li, X. Zeng, Z. Zhang, D. Zhang, Y. Long, and N. Wang, "Neural network-based in-loop filter for CLIC 2022," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, New Orleans, LA, USA, Jun. 2022, pp. 1773–1776, doi: 10.1109/CVPRW56347.2022.00189.

[14] C. Jia, S. Wang, X. Zhang, S. Wang, J. Liu, S. Pu, and S. Ma, "Content-aware convolutional neural network for in-loop filtering in high efficiency video coding," *IEEE Trans. Image Process.*, vol. 28, no. 7, pp. 3343–3356, Jul. 2019, doi: 10.1109/TIP.2019.2896489.

[15] 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, doi: 10.1109/TCSVT.2021.3089498.

[16] 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, doi: 10.1109/ICCV.2015.73.

[17] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *Computer Vision—ECCV 2014* (Lecture Notes in Computer Science), D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham, Switzerland: Springer, 2014, pp. 184–199, doi: 10.1007/978-3-319-10593-2_13.

[18] C. Li, L. Song, R. Xie, and W. Zhang, "CNN based post-processing to improve HEVC," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 4577–4580, doi: 10.1109/ICIP.2017.8297149.

[19] 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, doi: 10.1109/ICME46284.2020.9102912.

[20] F. Zhang, D. Ma, C. Feng, and D. R. Bull, "Video compression with CNN-based postprocessing," *IEEE Multimedia*, vol. 28, no. 4, pp. 74–83, Oct./Dec. 2021, doi: 10.1109/MMUL.2021.3052437.

[21] M. Wang, S. Wan, H. Gong, Y. Yu, and Y. Liu, "An integrated CNN-based post processing filter for intra frame in versatile video coding," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Nov. 2019, pp. 1573–1577, doi: 10.1109/APSIPAASC47483.2019.9023240.

[22] X. Meng, X. Deng, S. Zhu, and B. Zeng, "Enhancing quality for VVC compressed videos by jointly exploiting spatial details and temporal structure," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 1193–1197, doi: 10.1109/ICIP.2019.8804469.

[23] M. Lu, T. Chen, H. Liu, and Z. Ma, "Learned image restoration for VVC intra coding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops*, Jun. 2019, pp. 1–4.

[24] 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, doi: 10.1109/TMM.2021.3108943.

[25] *FFmpeg Documentation*. Accessed: Mar. 22, 2023. [Online]. Available: https://ffmpeg.org/ffmpeg.html

[26] A. Wieckowski, J. Brandenburg, T. Hinz, C. Bartnik, V. George, G. Hege, C. Helmrich, A. Henkel, C. Lehmann, C. Stoffers, I. Zupancic, B. Bross, and D. Marpe, "VVenC: An open and optimized VVC encoder implementation," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Jul. 2021, pp. 1–2, doi: 10.1109/ICMEW53276.2021.9455944.

[27] A. Wieckowski, G. Hege, C. Bartnik, C. Lehmann, C. Stoffers, B. Bross, and D. Marpe, "Towards a live software decoder implementation for the upcoming versatile video coding (VVC) codec," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2020, pp. 3124–3128, doi: 10.1109/ICIP40778.2020.9191199.

[28] E. Alshina, R.-L. Liao, S. Liu, and A. Segall, *JVET Common Test Conditions and Evaluation Procedures for Neural Network Based Video Coding Technology*, document JVET-AC2016-v1, Joint Video Experts Team (JVET), Jan. 2023.
[29] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.* Red Hook, NY, USA: Curran Associates, 2019, pp. 1–12. Accessed: Mar. 4, 2023. [Online]. Available: https://proceedings.neurips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html
[30] G. Bjontegaard, *Calculation of Average PSNR Differences Between RD-Curves*, document VCEG-M33, Video Coding Experts Group (VCEG), Apr. 2001.

TANNI DAS received the B.Sc. degree from the Department of Electrical and Electronics Engineering, University of Chittagong, Chittagong, Bangladesh, in 2017. She is currently pursuing the M.Sc. degree with the Department of AI-Software, Gachon University, South Korea. In 2022, she joined as a Graduate Research Assistant with Gachon University. Her current research interests include image and video coding and AI-based multimedia technology.

KIHO CHOI (Member, IEEE) received the B.S. (summa cum laude) and Ph.D. degrees from the Department of Communication and Electronics and Computer Engineering, Hanyang University, Seoul, South Korea, in 2008 and 2012, respectively. In 2013, he was a Research Associate and a Lecturer with Hanyang University, before joining Samsung Research, Seoul, in 2014. He has been a Technical Leader with the Department of Video Coding Standards with Samsung Research, since 2014. From 2021 to 2023, he was an Associate Professor with the AI-Software Department, Gachon University. Currently, he is an Associate Professor with the Department of Electronic Engineering, Kyung Hee University. He has authored more than 200 MPEG contribution articles and more than 50 journals and conference papers. His current research interests include image and video coding, multimedia data compression, multimedia streaming, and AI-based multimedia technology. Since 2009, he has been an active participant in standardization for multimedia. He has served several chairs of ad hoc groups in ISO/IEC MPEG and JVET committees.

JAEYOUNG CHOI (Member, IEEE) received the B.S. and M.S. degrees from the Department of Mathematics, Korea University, South Korea, in 2008 and 2013, respectively, and the Ph.D. degree from the Department of Electrical Engineering, KAIST, in 2018. From 2018 to 2020, he was an Assistant Professor with the Department of Automotive Engineering, Honam University, South Korea. Since 2020, he has been an Assistant Professor with the School of Computing, Gachon University, South Korea. His research interests include the intersection of applied mathematics and statistical inference, including social networks, wireless vehicular networks, and probabilistic graphical models.