Neural Computing and Applications (2022) 34:14135–14149
https://doi.org/10.1007/s00521-021-06491-9

S.I. : EMERGING TRENDS IN AI & ML

Deep learning-based video quality enhancement for the new versatile video coding

Soulef Bouaafia[1] · Randa Khemiri[1,2] · Seifeddine Messaoud[1] · Olfa Ben Ahmed[3] · Fatma Ezahra Sayadi[4]

Received: 8 February 2021 / Accepted: 30 August 2021 / Published online: 8 September 2021

Abstract

Multimedia IoT (M-IoT) is an emerging type of Internet of things (IoT) relaying multimedia data (images, videos, audio and speech, etc.). The rapid growth of M-IoT devices enables the creation of a massive volume of multimedia data with different characteristics and requirements. With the development of artificial intelligence (AI), AI-based multimedia IoT systems have been recently designed and deployed for various video-based services for contemporary daily life, like video surveillance with high definition (HD) and ultra-high definition (UHD) and mobile multimedia streaming. These new services need higher video quality in order to meet the quality of experience (QoE) required by the users. Versatile video coding (VVC) is the new video coding standard that achieves significant coding efficiency over its predecessor high-efficiency video coding (HEVC). Moreover, VVC can achieve up to 30% BD rate savings compared to HEVC. Inspired by the rapid advancements in deep learning, we propose in this paper a wide-activated squeeze-and-excitation deep convolutional neural network (WSE-DCNN) technique-based video quality enhancement for VVC. Therefore, we replace the conventional in-loop filtering in VVC by the proposed WSE-DCNN model that eliminates the compression artifacts in order to improve visual quality and hence increase the end user QoE. The obtained results prove that the proposed in-loop filtering technique achieves −2.85%, −8.89%, and −10.05% BD rate reduction for luma and both chroma components under random access configuration. Compared to the traditional CNN-based filtering approaches, the proposed WSE-DCNN-based in-loop filtering framework achieves efficient performance in terms of RD cost.

Keywords Video coding · Artificial intelligence · Multimedia IoT · VVC

1 Introduction

The growing multimedia portfolio, including big data processing, cloud computing, and the Internet of things (IoT) [1], has a direct impact on our lifestyle. Multimedia IoT (M-IoT) is considered as a major network technology enabling the interconnection and interaction between humans, health centers, industries, and objects like cameras, transport, and sensors [1, 2]. In addition, M-IoT systems combine the networking technologies for computer vision, image processing, and connectivity. Yet, they can be used in driving assistance, surveillance such as crime and fire detection, and remote sensing such as high-speed object tracking [3]. Real-world multimedia applications,

✉ Soulef Bouaafia
soulef.bouaafia@fsm.rnu.tn

Randa Khemiri
randa.khemiri@gmail.com

Seifeddine Messaoud
seifeddine.messaoud@fsm.rnu.tn

Olfa Ben Ahmed
olfa.ben.ahmed@univ-poitiers.fr

Fatma Ezahra Sayadi
sayadi_fatma@yahoo.fr

[1] Laboratory of Electronics and Microelectronics, Faculty of Sciences of Monastir, University of Monastir, Monastir, Tunisia

[2] Higher Institute of Computer Science and Multimedia of Gabes, University of Gabes, Gabes, Tunisia

[3] XLIM-CNRS, Bât SP2MI, University of Poitiers, 11 Bd Marie et Pierre Curie, 86962 Chasseneuil Cedex, France

[4] Laboratory of Networked Objects Control and Communication Systems, National Engineering School of Sousse, University of Sousse, Sousse, Tunisia

including smart industry 4.0 and agriculture 4.0, smart traffic monitoring, smart cities, smart homes, smart health, and smart environment with intelligent surveillance systems [3], are illustrated in Fig. 1. However, several issues such as interoperability, security, data size, reliability, storage, and computational capacity need to be well resolved to process multimedia data [4].

Compared to traditional IoT, M-IoT has powerful functionality such as fast and reliable data delivery. Therefore, it imposes high quality of service (QoS) requirements and demands efficient network architecture. In this context, quality of experience (QoE) represents the perspective of the end user's QoS. QoE can be depicted as objective or subjective. The users' objective QoE is difficult to measure and varies considerably according to the requirements of M-IoT devices (bigger memory, higher computational power, more power-hungry with higher bandwidth, etc.). However, service providers concern with the subjective QoE to evaluate the network mean opinion score (MOS) [5]. Multimedia data (audio, image, video, etc.) pose several challenges for transmitting, storing, and sharing data, especially their processing [6]. Furthermore, M-IoT processing requires efficient feature extraction, event processing, encoding/decoding, energy-efficient computing, QoS, and QoE [7].

As emerging technologies have rapidly evolved, multimedia services and video applications have grown tremendously. Higher image resolution (4K, 8K), especially for video games and monitoring tasks, is needed to satisfy the QoS specifications of end users. In traditional multimedia encoding methods, data are compressed only one time and decoded in every playing time. M-IoT devices are more concerned with uploading the data in uplink transmission. The latter poses challenges on computationally powered constrained M-IoT devices. Therefore, versatile video coding (VVC) which is a powerful multimedia encoding/decoding technique has been widely adopted. VVC [8] is the new generation video coding developed in July 2020, by the joint video experts team (JVET), as a successor of the high-efficiency video coding (HEVC) [9]. As the next standard for sophisticated video coding technology, VVC allows up to 30% for BD rate savings while maintaining the same quality as HEVC. Although VVC aims to maintain high-quality compressed video with additional encoding features, these are still compression artifacts that can lead to lower QoE. Hence, the QoE of VVC compressed video needs to be improved.

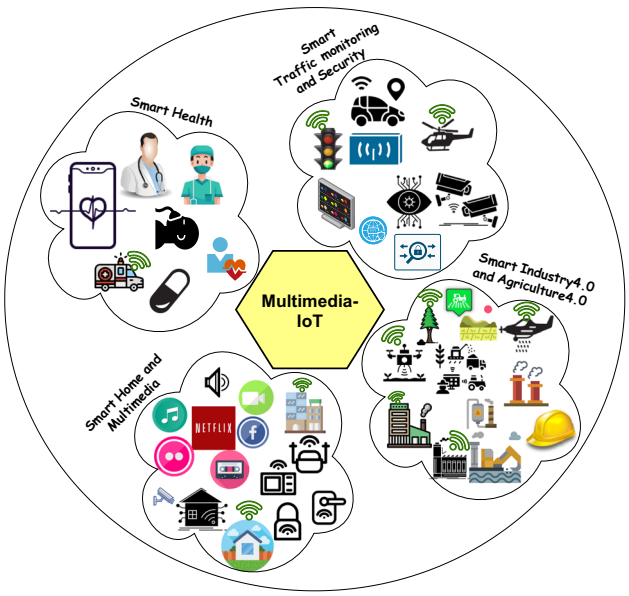

Fig. 1 M-IoT use cases

On the other hand, VVC adopts the block coding and the quantization structure; many different forms of distortion still exist, such as blocking artifacts, blurring, and ringing artifacts. The blocking artifacts affect the visual quality. While these distortions are permanent and cannot be removed entirely, special filters can be used to reduce them. For example, loop filters play an important role in reducing artifacts problems and in improving video and image qualities.

Unlike HEVC, in-loop filtering techniques [i.e., deblocking filter (DBF), sample adaptive offset (SAO), and adaptive loop filter (ALF)] are applied in the VVC standard. These filters remove the video compression artifacts and enhance the visual quality of the reconstructed video. Indeed, the DBF purpose is designed with the use of discontinuity-based smoothing filters to minimize artifacts along block boundaries [10, 11]. In order to reduce ringing artifacts through compensation, SAO is used as a filter added after DBF, which applies shifts to samples based on the encoder lookup table and analyzes signal amplitudes using a histogram [12]. ALF is the latest loop filtering considered as a new feature in VVC. ALF reduces distortions between reconstructed and original images [13]. Although these conventional in-loop filtering can relieve specific artifacts, it is difficult to overcome the complex distortion introduced by video compression. To meet this challenge, powerful deep learning approaches have been used. Among these techniques, convolutional neural network (CNN) is the most robust and efficient processing method for recognizing and analyzing images and videos [14–16].

Several CNN-based filtering approaches for HEVC and VVC standards have been proposed for video quality enhancement [17, 18]. These approaches using CNN-based in-loop filtering and post-processing are proposed to reduce visual artifacts and to achieve high performance. Indeed, regarding the challenges of 5G and M-IoT technologies, such as low latency cost, high speed rate, and high video

and image resolution quality, the VVC original loop filtering became insufficient to meet resolution requirement of M-IoT-based applications. To address these critical issues, QoE must be considered and improved in order to ensure QoS for end users [3].

In this context, we propose a deep CNN-based in-loop filtering approach, denoted as the wide-activated squeeze-and-excitation deep convolutional neural network (WSE-DCNN). The proposed approach provides new powerful in-loop filtering without exploiting traditional ones (DBF, SAO, ALF) for the VVC standard. Indeed, the main goal is to effectively remove compression artifacts and enhance the compressed video quality. The proposed method improves the QoE of end users. The contribution of this paper is summarized as follows:

- We propose a WSE-DCNN framework-based aware quality system for the M-IoT context.
- We implement the proposed scheme into VVC standard, which achieves coding gains accordingly for random access configuration.
- We adapt the M-IoT scenario-based smart city context in which QoE of video quality is improved.

The remainder of this paper is organized as follows: Sect. 2 presents the related work, and Sect. 3 introduces the proposed M-IoT scenario. Then, the proposed deep CNN-based in-loop filtering in VVC standard is defined in Sect. 4. Next, we evaluate the proposed method in Sect. 5. Finally, Sect. 6 concludes the paper and opens the same perspectives.

2 Related work

In this section, we start by briefly describing several existing works related to multimedia data computing in IoT for video coding. Then, we will present deep CNN-based in-loop filtering methods.

2.1 Video coding for M-IoT

M-IoT poses several challenges to identify data transmission methods that may have reduced latencies for real-time processing, while ensuring QoS, QoE, and flexible data sizes to meet bandwidth limitations and to reduce power consumption. To effectively reduce data transmission and improve video quality, video compression is the most interesting module to deal with in this field.

Video compression is therefore necessary for an efficient transmission of video data via band-limited Internet. This need was the most felt during the COVID-19 pandemic where data traffic is used for e-learning, video conferencing, and real-time surveillance. Lee et al. [19] proposed an encoding algorithm using HEVC for compressing high video quality with 4K and 8K resolutions. The fast proposed algorithm achieves better performances in terms of computational complexity and bit rate. In [20], an IoVT platform is developed which combines HEVC and H.264/advanced video coding (AVC) for reliable video streaming in real time. Meanwhile, in the mHealth context, video compression is considered as the key technology and therefore has been widely used for real-time medical video communications in a mHealth environment. With regards to healthcare application, Panayides et al. [21] studied VVC and AV1 (AOMedia Video 1) video encoding in mHealth video communication scenarios. In [22], a comparative study between HEVC and VVC was presented in the context of video telehealth systems. The obtained results prove that VVC requires a BD rate of up to 40%, with respect to a high quality in full high definition (FHD) video. On the other hand, Alarifi et al. [23] proposed a novel hybrid cryptosystem for secure streaming of HEVC in IoT multimedia applications.

However, these aforementioned methods are still limited in terms of QoE and QoS to be adapted to the new generation of wireless networks. The latter could usher in new immersive experiences, such as virtual reality (VR) and augmented reality (AR).

2.2 Deep learning-based video coding

Recently, deep learning (a branch of artificial intelligence) has seen great success in computer vision tasks [24, 25], especially for video encoding [26–28]. Indeed, deep neural networks have been adopted to improve coding tools, including intra- and inter-prediction, transformation, quantization, and loop filtering for HEVC and VVC standards. Regarding the HEVC, Bouaafia et al. [28] proposed a machine learning-based HEVC complexity reduction in the inter-prediction process. The proposed algorithm achieves good RD complexity performances. Additionally, for intra-coding context, a fast algorithm based on CNN to improve HEVC intra-coding performance is introduced in [29]. With regards to in-loop filtering in HEVC, Pan et al. [30] proposed an in-loop filtering using an enhanced convolutional neural network (ED-CNN) to replace DBF and SAO, in order to eliminate artifacts. The suggested scheme achieved 6.45% BD rate reduction and 0.238 dB PSNR gains. Variable-filter-size residue learning convolutional neural network (VRCNN) was proposed in [31] as a new technique for both DBF and SAO in intra-coding HEVC. The simulation results show that the proposed technique achieves a BD rate savings of 4.6%.

For VVC standard, Ma et al. [32] developed a new CNN model, MFRNet, as a way to improve loop-through filtering and post-processing. The proposed technique was

integrated into VVC test model to remove visual artifacts and enhance video quality. Furthermore, an in-loop filtering-based dense residual convolutional neural network (DRN) for VVC was proposed applied after DBF, and before SAO and ALF [18]. To reduce CU partition complexity, the fast intra-CU coding technique of H.266/VVC is implemented based on the enhanced DAG-SVM classifier model [33]. Experimental results show that the proposed model reaches 54.74% of encoding time. Therefore, Park et al. suggested a lightweight neural network (LNN) based on a fast decision algorithm to eliminate redundant VVC block partitioning [34]. The proposed model achieves a trade-off between encoding complexity and compression performance. However, these approaches do not take into account a QoE in VVC standard in an M-IoT context.

In this context, we propose an in-loop filtering based on wide-activated squeeze-and-excitation deep CNN (WSE-DCNN) approach to enhance VVC video quality and achieve coding gains.

3 Proposed M-IoT scenario-based architecture for multimedia data

Without loss of generality, we propose an M-IoT scenario in the context of smart city, as illustrated in Fig. 2. It consists of a set of M-IoT devices, like cameras and multimedia devices, that are capable to acquire multimedia contents from the real and physical word. After that, the sensed multimedia data are sent to the centralized cloud computing for processing, via the network layer, using different transmission technologies such as LP-WAN [35]. M-IoT devices are more concerned with uploading data in uplink transmission, which poses challenges for constrained computational M-IoT devices. Several metrics can be considered, in this step, like the delay, jitter, and packet loss rate. Our interest is shifting to central computing, such as M-IoT data compression and encoding/decoding.

After M-IoT data acquisition step, data are compressed once and decoded whenever played. Traditionally, video encoding/compression is achieved by utilizing spatial and temporal redundancies. In this context, video quality is considered as the potential challenge in the VVC standard, that must be improved in this phase, especially when the huge collected multimedia data are structured/unstructured, with high velocity, and with different resolutions. Therefore, the QoE, depending on the video quality performances, is denoted as the metric that should be maximized. Based on the modeling, given in [36], the QoE is modeled considering the bit rate (BR) as formulated in (1).

$$QoE_{BR} = a \times log(BR) + b \tag{1}$$

where a and b denote coefficients determined during the experiment. However, this parameter, just like the PSNR metric, will also be used for the proposed WSE-DCNN-based in-loop filtering to evaluate video quality.

4 WSE-DCNN-based in-loop filtering for VVC-based M-IoT

The proposed WSE-DCNN framework is integrated into VVC standard, which replaces the original VVC in-loop filter module, as shown in Fig. 3. The main purpose of this proposed approach is to enhance the visual quality of the reconstructed frame while maintaining coding gains. The rate distortion optimization (*RDO*) technique is applied in order to confirm whether or not the proposed loop filter based on WSE-DCNN is used at each coding unit (*CU*). The *RDO* metric is given then by Eq. (2).

$$J = D + \lambda R, \tag{2}$$

where D represents the distortion between the original and the reconstructed frame, R indicates the coding bits needed, and λ is the Lagrange multiplier controlling the trade-off between D and R. To avoid a reduction in *RDO* efficiency, the coding tree unit (*CTU*) level on/off control is applied. The frame level filtering would be shut off to prevent over-signal, if the enhancement quality is not worth to cost the signaled bits. For each *CTU*, the *CTU* control flag shall be enabled when *RDO* performance reaches a better quality of the filtered *CTU*, otherwise the flag will be disabled.

The concept of the proposed architecture is illustrated in Fig. 4. The proposed architecture is shared by luma (Y) and two chroma (U and V), so the three components will be filtered simultaneously. The proposed WSE-DCNN model consists of six inputs, three denoting the *YUV* reconstructed and the three others include the quantization parameter *QP* and the coding unit (CU) for luminance and chrominance. Meanwhile, these inputs are first normalized to provide better convergence in the learning phase and then fed to a WSE-DCNN-based in-loop filtering. Hence, the three (*Y*/*U*/*V*) reconstructions are normalized to [0, 1] based on the highest bit depth value. This implies that the normalized values ($P'(x, y)$) are achieved by Formula (3).

$$P''(x, y) = \frac{P'(x, y)}{1 << B - 1}, \quad x = 1, \ldots, W, y = 1, \ldots, H \tag{3}$$

where the bit depth is denoted by B, $P''(x, y)$ is the normalized value in normalized *Y*/*U*/*V* at (x, y), and W and H are the width and the height of the reconstructed frame, respectively.

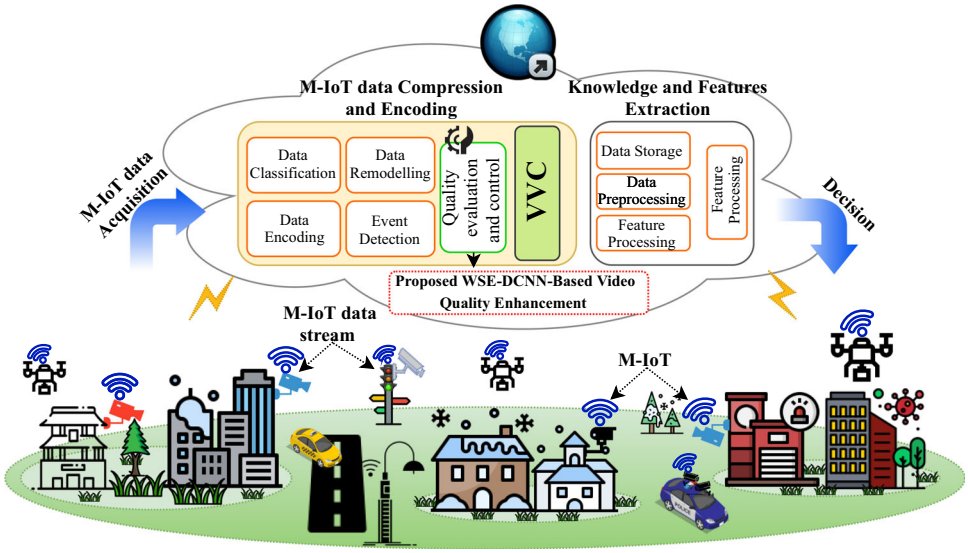

Fig. 2 M-IoT scenario-based centralized video quality enhancement

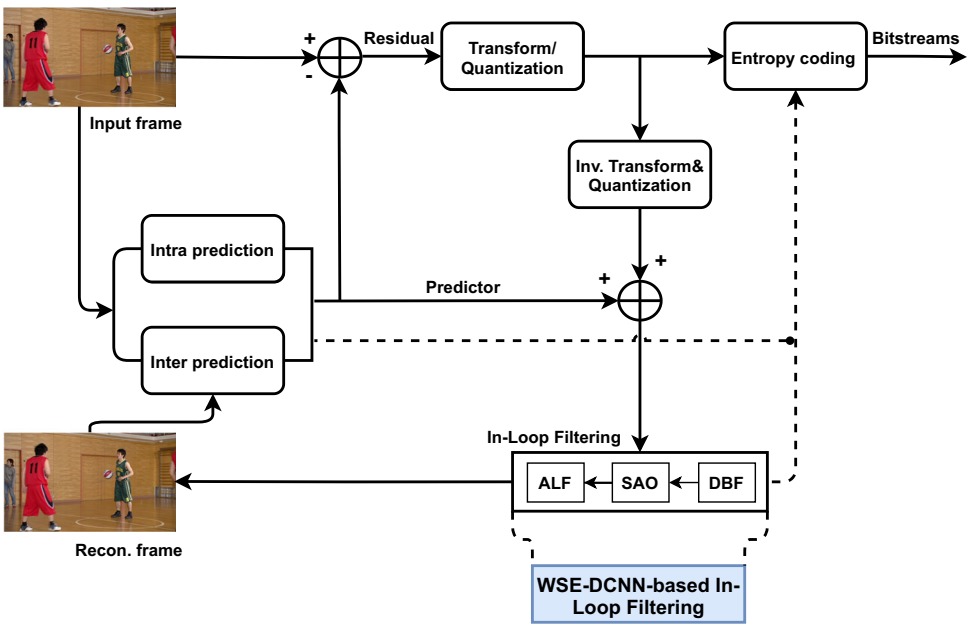

Fig. 3 Proposed VVC standard framework

Various quantization parameters (QPs) contribute to a variety of reconstructed video quality. This makes it easier to use a single set of parameters to fit reconstructions with different qualities. *QP* should be normalized to *QPmap* following Formula (4).

$$QPmap(x,y) = \frac{QP}{63}, \quad x = 1,\ldots,W, y = 1,\ldots,H \tag{4}$$

Regarding the other inputs, there are *CU* partition of the luma (*Y*) and chroma (*UV*) components. Since the blocking artifacts are mainly caused by CU block partition. The *CU* block partition is converted into coding unit maps (*CUmaps*) and normalized, and then it is considered as an input to the network. For example, for each *CU* in each frame, the boundary position is filled with two and the other positions are filled with one. However, two *CUmaps* can be obtained, one as $Y - CUmap$ for luma and the other denoted by $UV - CUmap$ for chroma.

As shown in Fig. 4, the processing of WSE-DCNN-based in-loop filtering has three levels. At the first level, the three components of *YUV* are processed through WSE blocks, and each one of them is fused with its corresponding *CUmap*. In addition, *CUmap* would be multiplied by its own corresponding channel before being concatenated with feature maps. In the second level, the feature maps of different channels are connected together and then processed by several WSE blocks. At this level, the *QPmap*

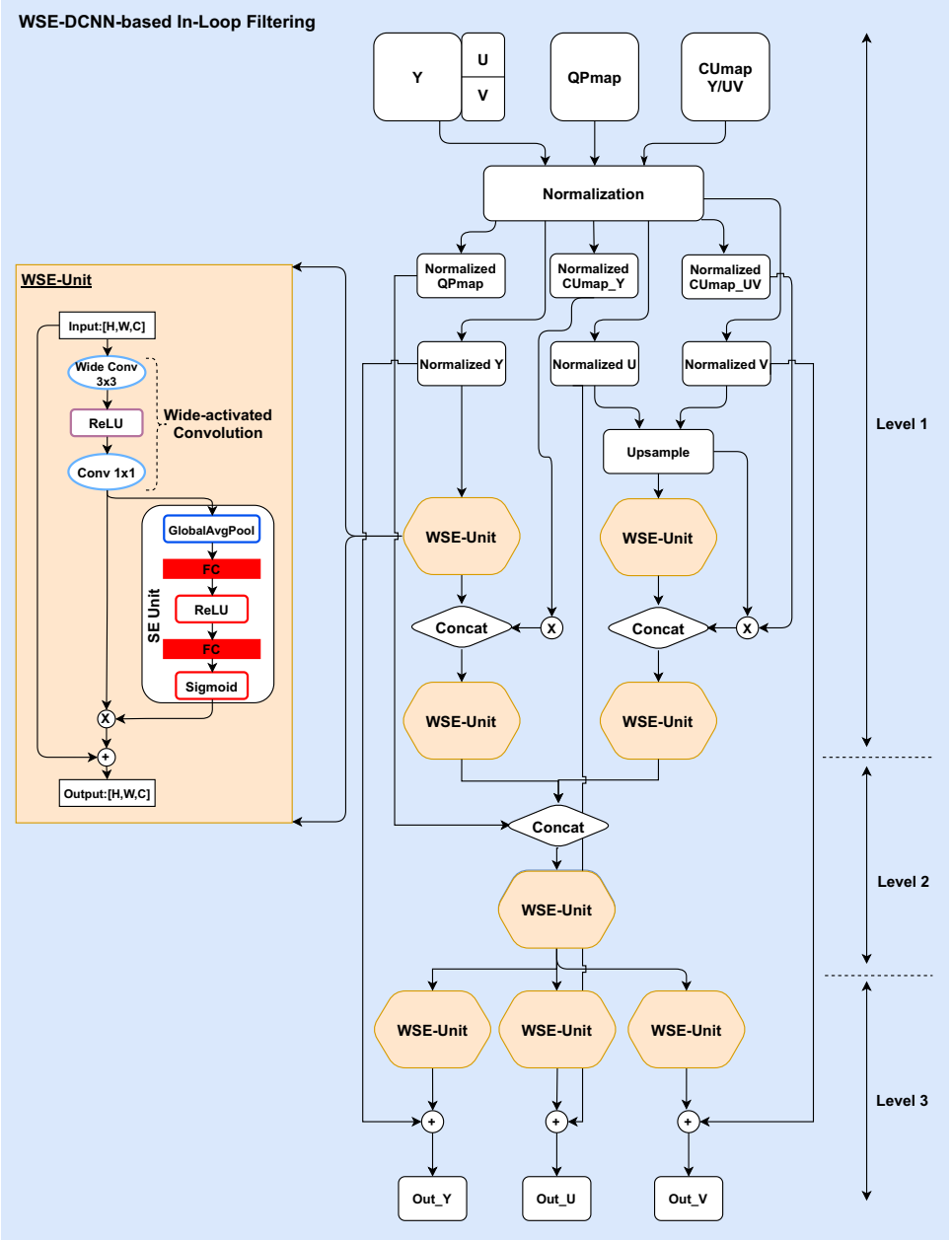

Fig. 4 WSE-DCNN architecture

is also concatenated. Finally, in the last level, the three channels are processed separately again to generate the final residual image. Then, the original input will be implemented as a residual CNN. The WSE module is considered to be the basic unit of the WSE-DCNN-based in-loop filtering proposed in the VVC standard and shown in Fig. 4. Additionally, this basic unit is composed of the wide-activated convolution [37] and the squeeze-and-excitation (*SE*) [38] operation. The wide-activated convolution performs very well in super-resolution and noise reduction tasks. It consists of a wide 3×3 convolution followed by ReLU (rectified linear unit) [39, 40] activation function and a narrow 1×1 convolution. Next comes the *SE* operation which is the most technical operation used to weight each convolutional layer. It can use the complex relationship between different channels and generates a weighting factor for each channel.

The WSE unit consists of the following phases as depicted in Algorithm 1, given a feature map X with shape $H \times W \times C$, where C means channel amounts:

Algorithm 1: WSE-Unit

Input: $X \in \{H, W, C\}$
Output: $Y \in \{H, W, C\}$

1 **for** *number of Epochs* **do**
2 **while** *True* **do**
3 **Wide-activated Conv-Function(X):**
4 $Y_1 = ReLU(W_1 X + b_1)$
5 $Y_2 = W_2 Y_1 + b_2$
6 **return** (Y_2)
7 **Squeeze-and-Excit-Function**(Y_2)**:**
8 **while** *True* **do**
9 ***Call-Squeeze-Operation***(Y_2)***:***
10 $Y_3(k) = \frac{1}{HXW} \sum_{i=1}^{H} \sum_{j=1}^{W} Y_2(i, j, k).$
11 **return**(Y_3)
12 ***Call-Excitation-Operation***(Y_3)***:***
13 $Y_4 = ReLU(W_4 Y_3 + b_4).$
14 $Y_5 = \sigma(W_5 Y_4 + b_5).$
15 **return**(Y_5)
16 **END while**
17 **return**(Y_5)
18 **WSE Function**(Y_2, Y_5)**:**
19 $Y_6(i, j, k) = Y_2(i, j, k) \times Y_5(k),$
20 $\forall i \in \{1, ..., H\}, \forall j \in \{1, ..., W\},$
21 $\forall k \in \{1, ..., C\}.$
22 **return**(Y_6)
23 **END while**
24 **END for**

- A wide 3×3 convolution followed by ReLU and a convolution layer with kernel size is 1×1. Given Y_1 is the channel defined in Algorithm 1 line 4 and Y_2 is the output of the second convolution layer given in line 5.
- Each channel obtains a value according to the squeeze operation using global average pooling (GAP) $Y_3(k)$ as shown in Algorithm 1 line 10.
- The excitation operation is described by two fully connected layers followed by ReLU and sigmoid (σ) activation functions, respectively. As shown in Algorithm 1 line 13, Y_4 is the first fully connected layer followed by ReLU, which is refined by a certain ratio r. Then, the second fully connected layer is followed by the sigmoid activation function which is denoted by Y_5 in line 14, and it gives each channel a smoothing gating ratio in the range of [0,1].
- According to WSE function, each Y_2 channel is multiplied by the gating ratio r, as defined in line 19.
- Finally, when the number of input is equals to the output channels C, a skip connection will be added directly from input to output to learn the residue. Otherwise, there is no skipped connection.

5 Experimental Results

In this section, we evaluated the performances, in terms of RD performance, and QoE of the proposed WSE-DCNN-based in-loop filtering scheme in VVC standard. Then, a comparison to the state of the art is made.

5.1 Dataset collection

In this work, we exploited the public large video dataset BVI-DVC [41], especially developed for training the deep video compression methods. According to the work cited in [41], all selected sequences are progressive-scanned at a spatial resolution of 3840×2160, with frame rates ranging from 24 fps to 120 fps, a bit depth of 10 bit, and in YCbCr 4:2:0 format. All of them are truncated to 64 frames without scene cuts, using the segmentation method described in [42]. To further increase data diversity and provide data augmentation, the 200 video clips were spatially down-sampled to 1920×1080, 960×540, and 480×270 using a Lanczos third-order filter. The BVI-DVC dataset includes 800 video sequences with different contents at four different resolutions. Table 1 summarizes the key features of BVI-DVC video training dataset used in this study.

In this context, we selected 80% video sequences for the training model and 20% for validation from the BVI-DVC video training dataset. These sequences are compressed by VVC reference software (VTM-4.0) [43] with QP values (22, 27, 32, 37) under random access configuration. For each QP, the reconstruction video images, including luma and chroma components, and its corresponding ground truth are divided into 64×64 patches, which were selected in a random order.

5.2 Deep model training, testing, and evaluation

The proposed deep learning model is trained offline in a supervised learning manner. During training phase, the TensorFlow GPU [44] is used as a deep learning framework to train the proposed model. The training parameters used in our experiments are summarized as follows: The batch size is set to 128, the training epochs to 200, the learning rate to 0.001, and weight decay of 0.1 for every 50 epochs. The Adam [45] optimizer is used to train our deep model. The training platform uses windows 10 OS with Intel®core TM i7-3770 @3.4 GHz CPU and 16 GB RAM and an NVIDIA GeForce RTX 2070 GPU.

Table 1 Key features of BVI-DVC video training database [41]

Features	BVI-DVC [41]
Image or videos	Video
Sequences number	800
Images number in each video	64
Max resolution	2160 p
Min resolution	270 p
Bit depth	10
Various textures	Yes

The mean square error (MSE) is applied as a loss function between the ground truth image and the reconstructed image [35]. Equation (5) defines the MSE loss function.

$$L(\theta) = \frac{1}{N}\sum_{i=1}^{N} ||F(Y_i, \theta) - X_i||_2^2 \tag{5}$$

Let X_i be the ground truth of the proposed model, where $i \in \{1, \ldots, N\}$. The output of the WSE-DCNN model is denoted by $F(\cdot)$, where Y_i represents the compressed images, $i \in \{1, \ldots, N\}$, and θ is the parameter set of the proposed framework. The loss function has indeed converged on a minimum value; it means that our model is well trained. To prove the efficiency of the proposed WSE-DCNN network, Fig. 5 shows the MSE loss and the validation peak signal-to-noise ratio (PSNR) during training process. The PSNR is defined by Eq. (6) [46]. As we can see, the MSE loss function performs well in terms of the convergence's performance.

$$PSNR = 10 \times log \frac{(2^B - 1)^2}{MSE} \tag{6}$$

where B is the number of bits per sample of the video sequence and the MSE is defined in Eq. (5).

In the testing phase, our proposed WSE-DCNN model is integrated into VVC standard to replace the traditional in-loop filtering method. All simulations are tested under the VVC JVET common test conditions (*CTC*) [47] using random access configuration at *QP* values (22, 27, 32, 37). The (VTM-4.0) with traditional filters enabled is used in our experiments. From VVC CTC, 17 test sequences were used for performance evaluation, including class A1 (3840×2160), class A2 (3840×2160), class B (1920×1080), class C (832×480), and class D (416×240). To evaluate the coding performance of the proposed model, Bjontegaard delta bit rate (BD rate) [48] is applied as an assessment metric.

5.3 WSE-DCNN evaluation

The RD performance results of the proposed model compared to the original VVC standard are shown in Table 1. Columns *Y*, *U*, and *V* in the table show the BD rate of *Y*, *U*, and *V* components, respectively. Ratios of the encoding and decoding time are denoted by T_{enc} and T_{dec} of the proposed model compared to the original one. The encoding time is defined by Eq. (7), where the coding complexity of the proposed method is defined by T_{Pro} and the coding complexity of the original VVC is denoted by T_{Orig}.

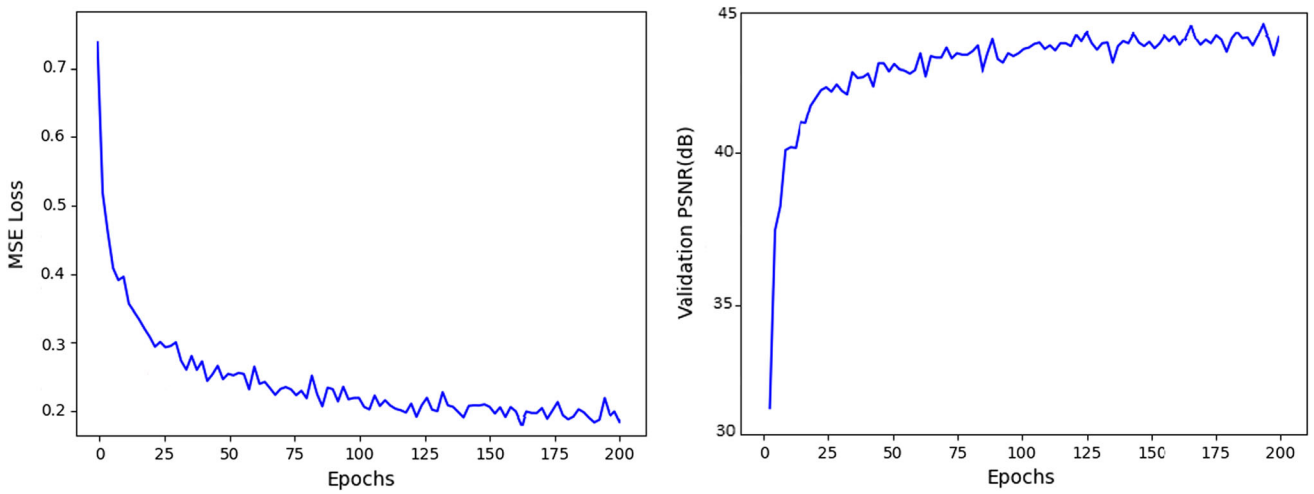

Fig. 5 Training MSE loss and validation PSNR

$$T = \frac{T_{Pro}}{T_{Orig}} \times 100\% \tag{7}$$

As shown in Table 2, the proposed scheme achieves better mean coding gains when integrated into VVC with 2.85% BD rate savings for luma Y component under random access configuration, while achieving 8.89% and 10.05% BD rate reduction for both chroma U and V components. The proposed scheme offers significant RD compression performance mainly in U and V chrominance for all test sequences. It is also clear that the coding performance differs for several sequences. This means that the proposed model is impacted by video sequence information. Moreover, the proposed model performs well in terms of coding gains for high motion or rich texture video sequences, as well as Tango2, DaylightRoad2, Kimono2, RaceHorses, etc. In summary, the proposed technique outperforms better than the VVC with traditional in-loop filtering algorithm in terms of RD performance.

Regarding complexity reduction, the time differences between the proposed VVC algorithm and the original VVC standard for encoding and decoding are summarized in Table 2. NVIDIA GeForce RTX 2070 GPU is used to measure the encoding and decoding time of the proposed

Table 2 Performance evaluation of the proposed model under random access configuration

Class	Sequences	Y (%)	U (%)	V (%)	T_{enc} (%)	T_{dec} (%)
Class A1	Tango2	− 2.89	− 10.02	− 11.35	147	939
4K	Campfire	− 1.22	− 2.75	− 10.28	119	1537
Class A2	CatRobot1	− 1.89	− 10.76	− 8.03	151	975
4K	DaylightRoad2	− 1.47	− 12.36	− 2.55	149	875
Class B	Kimono2	− 0.51	− 8.13	− 20.63	82	1524
1080p	ParkScene	− 4.18	− 9.25	− 12.94	125	1947
	Cactus	− 2.36	− 12.27	− 9.70	117	1154
	BasketballDrive	− 2.53	− 4.83	− 7.82	107	1981
	BQTerrace	0.11	− 2.88	0.63	116	1619
Class C	BasketballDrill	− 3.84	− 7.01	− 9.97	137	1312
WVGA	BQMall	− 3.89	− 11.48	− 10.92	118	1137
	PartyScene	− 4.65	− 9.69	− 9.63	115	1112
	RaceHorses	− 1.35	− 10.70	− 13.66	94	1079
Class D	BasketballPass	− 3.40	− 8.21	− 7.79	125	4628
WQVGA	BQSquare	− 5.27	− 4.39	− 11.44	137	2001
	BlowingBubbles	− 4.15	− 8.52	− 5.19	115	2045
	RaceHorses	− 5.08	− 18.04	− 19.74	114	2164
Overall		**− 2.85**	**− 8.89**	**− 10.05**	**122**	**1648.76**

Bold values indicate the mean of all performance values for all classes

filtering technique. From the table, we can observe that on average the encoding time overhead is 122% (for all test sequences: class A1 to class D), while the decoding time overhead is 1648.76% compared to the original VVC algorithm under random access configuration. The proposed scheme greatly influences the decoding time because of the forward operation in network and CPU-GPU memory copy operation. Therefore, we note that the proposed model achieves a little increase in encoding time compared to the original VVC algorithm.

To demonstrate the effectiveness of our proposed filtering model integrated into the VVC standard, PSNR is also used as a quality measure, which is calculated by the following equation [46]:

$$PSNR_{YUV} = \frac{6 \times PSNR_Y + PSNR_U + PSNR_V}{8} \quad (8)$$

The BQSquare video sequence encoded with *QP* equal to 37 under random access configuration is deployed in order to show the subjective visual quality and to further verify the effectiveness of the proposed model. Figure 6 shows the comparison of video subjective quality. It is clear that frame details are blurry when being compressed by the original VVC standard, but become clearer after being filtered by the proposed model. In fact, the proposed model effectively eliminates blocking artifacts as well as ringing artifacts and blurring, which improves visual quality, compared to the VVC standard with/without traditional in-loop filtering. Therefore, we compare the QoE variation with respect to bit rate of the proposed technique with the original VVC for class A1 to class D using random access configuration at four QPs, as shown in Fig. 7. It is remarkable that the suggested technique meets the QoE requirements of the end users, especially in high-resolution video sequences, as well as in class A1, class A2, and class B.

We also compared the proposed approach with other filtering models based on CNN network. Table 3 shows the comparison of encoding performance with other approaches [18, 49–51] in terms of reducing RD complexity under random access using VVC CTC. In [18], the authors proposed an in-loop filter algorithm-based dense residual convolutional neural network (DRN) to improve the reconstructed video quality. This network is integrated after *DF*, and before *SAO*, and *ALF* into VVC VTM-4.0 test model. This model is trained using the DIV2K dataset [52]. Moreover, a CNN-based in-loop filter algorithm is proposed for both intra- and inter-pictures placed before *ALF* with *DBF* and *SAO* are disabled [49]. This method is implemented into VVC VTM-3.0 standard [49]. Yet, in [50], the authors proposed a CNN-based loop filter placed in all traditional filters in VVC which is trained based on the DIV2K [52] dataset and implemented in VTM-5.0. In addition, Huang et al. [51] proposed a novel multi-gradient convolutional neural network-based in-loop filter for VVC to replace the original DBF and SAO filters. This network is trained based on the DIV2K dataset [52] and implemented in VTM-3.0.

As shown in Table 3, the proposed WSE-DCNN framework integrated into VVC standard achieves best *RD* performance for both luminance and two chrominance for

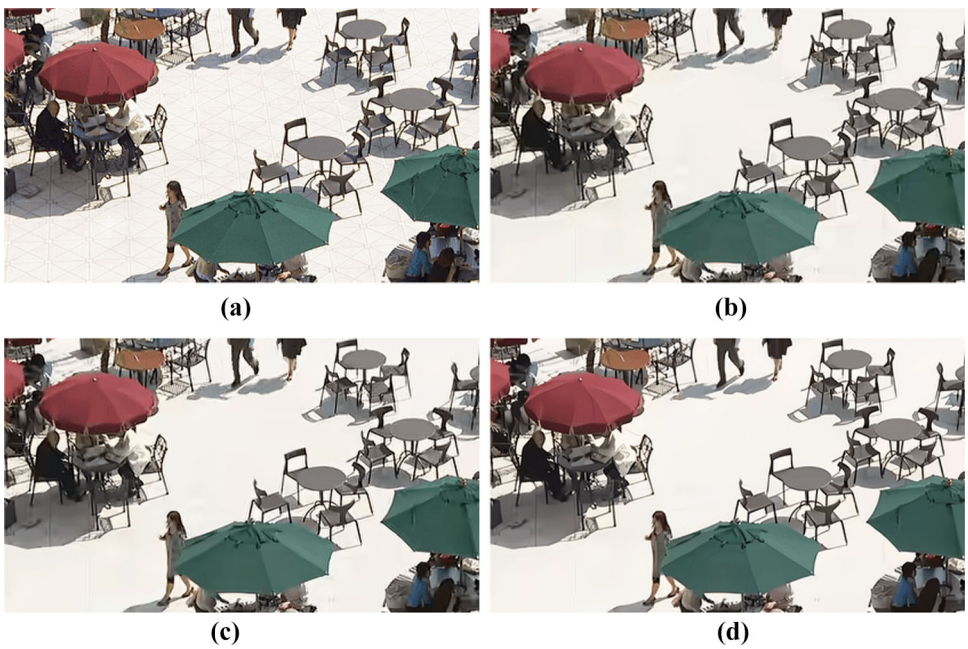

Fig. 6 Ablation study. Subjective visual quality comparison (the 12th frame of BQSquare with $QP = 37$: **a** original; **b** VVC without in-loop filtering ($PSNR = 31.17$ dB); **c** VVC ($PSNR = 31.37$ dB); **d** VVC-based proposed model ($PSNR = 31.68$ dB)

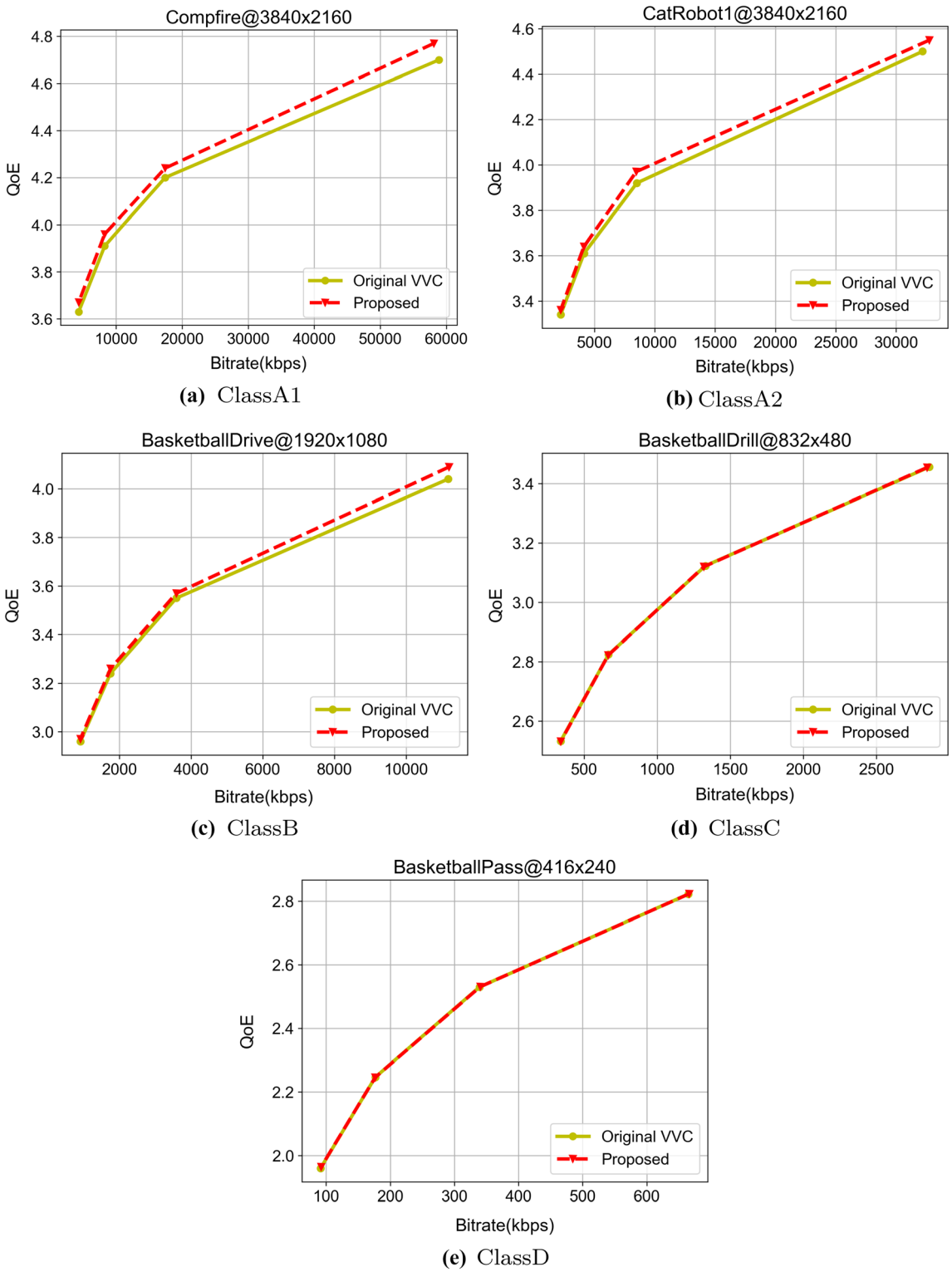

Fig. 7 Comparison of QoE variation with respect to bit rate

Table 3 Coding performance comparison with other approaches

Class	Schemes	Y (%)	U (%)	V (%)	T_{enc} (%)	T_{dec} (%)
Class A1	[18]	− 1.27	− 3.38	− 5.10	106	6967
4K	[49]	0.87	0.12	0.22	149	81,711
	[50]	0.18	0.63	− 2.95	138	123,657
	[51]	− 1.10	− 0.30	0.31	–	–
	Proposed model	**− 2.05**	**− 6.38**	**− 10.81**	**133**	**1238**
Class A2	[18]	− 2.21	− 5.74	− 2.88	106	6435
4K	[49]	− 1.12	− 0.52	− 2.11	142	81,263
	[50]	− 0.98	7.16	− 3.34	138	121,571
	[51]	− 1.94	0.89	− 2.24	–	–
	Proposed model	**− 1.68**	**− 11.56**	**− 5.29**	**150**	**925**
Class B	[18]	− 1.13	− 4.73	− 4.55	106	7011
1080p	[49]	− 0.83	− 0.47	− 1.20	143	69,595
	[50]	0.64	− 4.16	− 3.67	149	154,962
	[51]	− 2.51	− 1.28	− 0.89	–	–
	Proposed model	**− 1.89**	**− 7.47**	**− 10.09**	**109**	**1645**
Class C	[18]	− 1.39	− 3.63	− 4.36	106	8110
WVGA	[49]	− 1.76	− 3.64	− 6.80	122	46,645
	[50]	− 1.17	− 4.38	− 1.61	129	122,434
	[51]	− 4.03	− 4.17	− 5.62	–	–
	Proposed model	**− 3.43**	**− 9.72**	**− 11.05**	**116**	**1160**
Class D	[18]	− 1.39	− 1.96	− 3.08	105	4217
WQVGA	[49]	− 2.95	− 3.27	− 7.35	123	32,155
	[50]	− 3.13	− 6.26	− 5.15	122	104,265
	[51]	− 5.33	− 4.11	− 5.83	–	–
	Proposed model	**− 4.47**	**− 9.79**	**− 11.04**	**122**	**2709**

all test sequences from class A1 to class D, as compared to previous proposed approaches. These results imply that the proposed model performs well in terms of both objective and subjective visual qualities. Regarding the computational complexity, the proposed method outperforms other approaches in terms of encoding time for class A2. Compared with related works in [18, 49, 50], the methods proposed exceed our proposed scheme in terms of complexity reduction. In conclusion, the suggested method leads to efficient performance in almost all test sequences in terms of RD performance, demonstrating the efficiency and universality of the WSE-DCNN solution compared to other methods, whereas the computational complexity of VVC standard is still limited.

For further evaluation, we have provided *RD* performance curves of the suggested model-based in-loop filtering versus the other three methods under random access configuration with four *QPs* for class A1 to class D. Figure 8 shows the comparison in terms of *RD* performance (PSNR based on bit rate). By comparing the associated approaches, we can conclude that the proposed filtering model considerably improves the *RD* performance of the VVC standard. Our proposed in-loop filtering model works well especially in high-resolution video sequences, as well as in class A1, class A2, and class B.

6 Conclusion

In this paper, we proposed a deep learning algorithm-based VVC standard to enhance visual video quality while improving the user's QoE. The proposed WSE-DCNN framework is implemented into VVC standard to replace in-loop filtering in order to alleviate the coding artifacts, such as ringing, blocking, and blurring. The proposed VVC filtering technique is used in the M-IoT scenario-based smart city context to contribute to the centralized cloud that attempts to meet the required user's video quality. Compared to the traditional VVC-based filters, simulation results prove that the proposed framework achieves the best compression performance in terms of objective and subjective quality, with a BD rate savings about −2.85%, −8.89%, and −10.05% for Y, U, and V components, respectively. Therefore, this has proven the effectiveness of the proposed technique for video quality enhancement. Future works include the improvement in the VVC

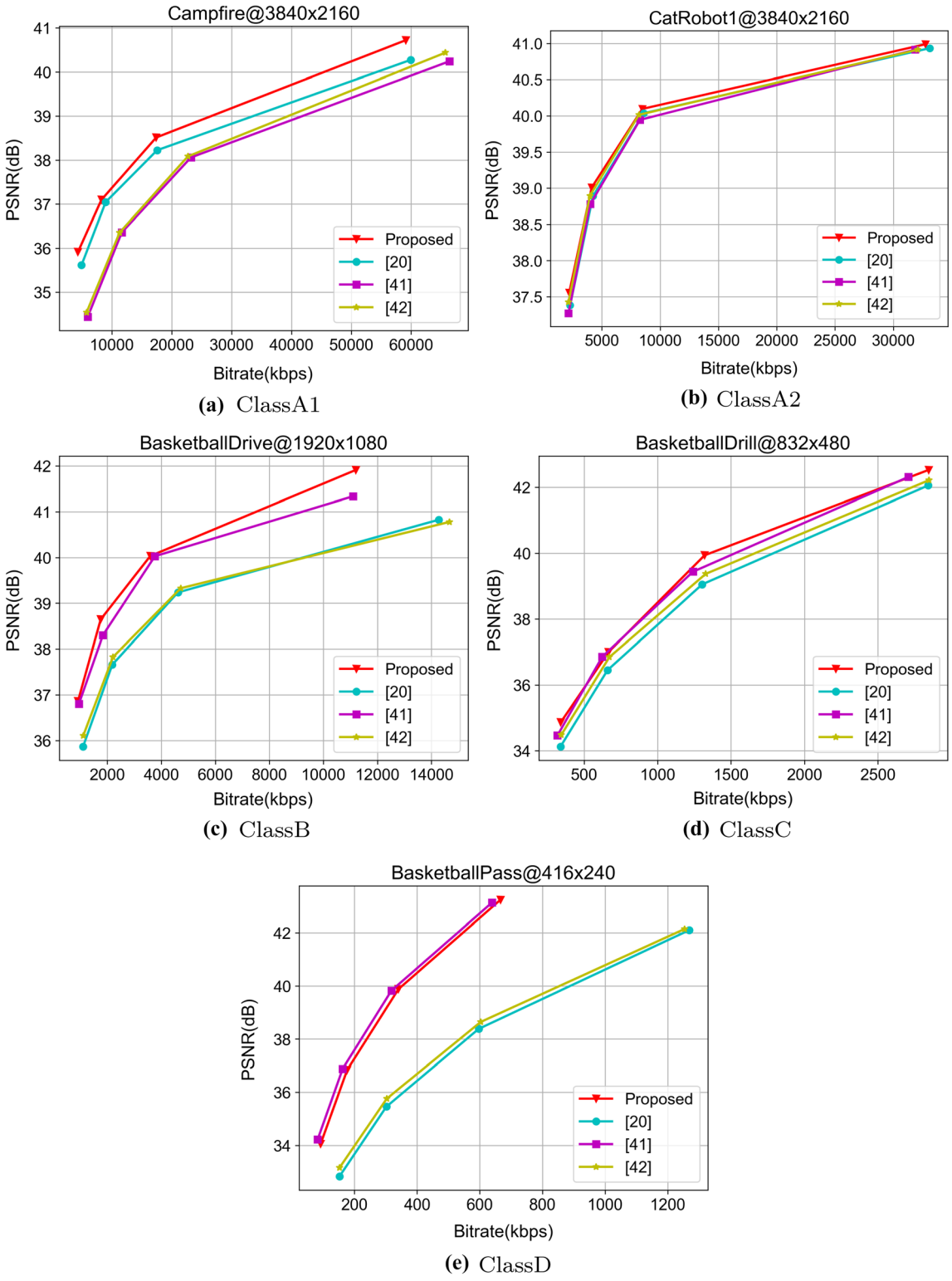

Fig. 8 RD performance curves of the proposed model compared to the other three approaches

computational complexity (time encoding and time decoding) [23].

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Messaoud S, Bradai A, Bukhari SHR, Qung PTA, Ahmed OB, Atri M (2020) A survey on machine learning in internet of things: algorithms, strategies, and applications. Internet of Things, 100314
2. Cao Y, Jiang T, Han Z (2016) A survey of emerging M2M systems: context, task, and objective. IEEE Internet of Things J 3(6):1246–1258
3. Nauman A, Qadri YA, Amjad M, Zikria YB, Afzal MK, Kim SW (2020) Multimedia internet of things: a comprehensive survey. IEEE Access 8:8202–8250
4. Zikria YB, Kim SW, Hahm O, Afzal MK, Aalsalem MY (2019) Internet of Things (IoT) operating systems management: opportunities, challenges, and solution. Multidisciplinary Digital Publishing Institute
5. Amjad M, Rehmani MH, Mao S (2018) Wireless multimedia cognitive radio networks: a comprehensive survey. IEEE Commun Surv Tutor 20(2):1056–1103
6. Marjani M, Nasaruddin F, Gani A, Karim A, Hashem IAT, Siddiqa A, Yaqoob I (2017) Big IoT data analytics: architecture, opportunities, and open research challenges. IEEE Access 5:5247–5261
7. Kumari A, Tanwar S, Tyagi S, Kumar N, Maasberg M, Choo K-KR (2018) Multimedia big data computing and Internet of Things applications: a taxonomy and process model. J Netw Comput Appl 124:169–195
8. Bross B, Chen J, Liu S (2019) Versatile video coding (Draft 4) JVET-M1001. In: 13th Meeting of the joint video exploration team (JVET), Marrakech, pp 9–18
9. Wien M (2015) High efficiency video coding. Coding Tools Specif 24
10. Ichigaya A, Iwamura S, Nemoto S (2018) Syntax and semantics changes of luma adaptive deblocking filter. In: The JVET meeting. Macao, China: ITU-T, ISO/IEC, number JVET-L0414
11. Kotra Meher A, Esenlik S, Wang B, Gao H, Alshina E (2019) Non-CE5: chroma QP derivation fix for deblocking filter. In: The JVET meeting. Geneva, Switzerland: ITU-T, ISO/IEC, number JVET-P1001
12. Browne A, Sharman K, Keating S (2020) SAO modification for 12-bit. In: The JVET meeting. Brussels, Belgium: ITU-T, ISO/IEC, number JVET-Q0441
13. Hu N, Seregin V, Karczewicz M (2019) Non-CE5: spec fix for ALF filter and transpose index calculation. In: The JVET meeting. Geneva, Switzerland: ITU-T, ISO/IEC, number JVET-Q0665
14. Krizhevsky A, Sutskever I, Hinton GE (2017) ImageNet classification with deep convolutional neural networks. Commun ACM 60(6):84–90
15. Mohan K, Seal A, Krejcar O, Yazidi A (2021) FER-net: facial expression recognition using deep neural net. Neural Comput Appl 1-12
16. Ben-Ahmed O, Huet B (2018) Deep multimodal features for movie genre and interestingness prediction. In: 2018 International conference on content-based multimedia indexing (CBMI). IEEE, pp 1–6
17. Jia W, Li L, Li Z, Liu S (2019) Residue guided loop filter for HEVC post processing. arXiv preprint arXiv:190712681
18. Chen S, Chen Z, Wang Y, Liu S (2020) In-Loop filter with dense residual convolutional neural network for VVC. In: 2020 IEEE conference on multimedia information processing and retrieval (MIPR). IEEE, pp 149–152
19. Lee J-H, Jang K-S, Kim B-G, Jeong S, Choi JS (2015) Fast video encoding algorithm for the internet of things environment based on high efficiency video coding. Int J Distrib Sens Netw 11(11):146067
20. Sammoud A, Kumar A, Bayoumi M, Elarabi T (2017) Real-time streaming challenges in internet of video things (IoVT). In: 2017 IEEE international symposium on circuits and systems (ISCAS). IEEE, pp 1–4
21. Panayides AS, Pattichis MS, Pantziaris M, Constantinides AG, Pattichis CS (2020) The battle of the video codecs in the healthcare domain-a comparative performance evaluation study leveraging VVC and AV1. IEEE Access 8:11469–11481
22. Usman MA, Usman MR, Naqvi RA, Mcphilips B, Romeika C, Cunliffe, D et al (2021) Suitability of VVC and HEVC for video telehealth systems. CMC-Comput Mater Continua 67(1):529–547
23. Alarifi A, Sankar S, Altameem T, Jithin KC, Amoon M, El-Shafai W (2020) A novel hybrid cryptosystem for secure streaming of high efficiency H. 265 compressed videos in IoT multimedia applications. IEEE Access 8:128548–128573
24. Sahu G, Seal A, Krejcar O, Yazidi A (2021) Single image dehazing using a new color channel. J Vis Commun Image Represent 74:103008
25. Sengupta A, Seal A, Panigrahy C, Krejcar O, Yazidi A (2020) Edge information based image fusion metrics using fractional order differentiation and sigmoidal functions. IEEE Access 8:88385–88398
26. Bouaafia S, Khemiri R, Sayadi FE, Atri M, Liouane NA (2020) Deep CNN-LSTM framework for fast video coding. International conference on image and signal processing. Springer, Berlin, pp 205–212
27. Bouaafia S, Khemiri R, Maraoui A, Sayadi FE (2021) CNN-LSTM learning approach-based complexity reduction for high-efficiency video coding standard. Sci Program
28. Bouaafia S, Khemiri R, Sayadi FE, Atri M (2020) Fast CU partition-based machine learning approach for reducing HEVC complexity. J Real-Time Image Process 17(1):185–196
29. Yeh C-H, Zhang Z-T, Chen M-J, Lin C-Y (2018) HEVC intra frame coding based on convolutional neural network. IEEE Access 6:50087–50095
30. Pan Z, Yi X, Zhang Y, Jeon B, Kwong S (2020) Efficient in-loop filtering based on enhanced deep convolutional neural networks for HEVC. IEEE Trans Image Process 29:5352–5366
31. Dai Y, Liu D, Wu FA (2017) convolutional neural network approach for post-processing in HEVC intra coding. International conference on multimedia modeling. Springer, Berlin, pp 28–39
32. Ma D, Zhang F, Bull D (2020) MFRNet: a new CNN architecture for post-processing and in-loop filtering. IEEE J Sel Top Signal Process
33. Zhang Q, Wang Y, Huang L, Jiang B, Wang X (2020) Fast CU partition decision for H. 266/VVC based on the improved DAG-SVM classifier model. Multimed Syst 1–14
34. Park S-H, Kang J (2020) Fast multi-type tree partitioning for versatile video coding using a lightweight neural network. IEEE Trans Multimed
35. Messaoud S, Bradai A, Ahmed OB, Quang P, Atri M, Hossain MS (2020) Deep federated Q-learning-based network slicing for industrial IoT. IEEE Trans Ind Inform

36. Pal D, Vanijja V (2017) A no-reference modular video quality prediction model for H. 265/HEVC and VP9 codecs on a mobile device. Adv Multimed 2017
37. Yu J, Fan Y, Yang J, Xu N, Wang Z, Wang X, Huang T (2018) Wide activation for efficient and accurate image super-resolution. arXiv preprint arXiv:180808718
38. Hu J, Shen L, Sun G (2018) Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7132–7141
39. Nair V, Hinton GE (2010) Rectified linear units improve restricted Boltzmann machines. Icml
40. KiseI'ák J, Lu Y, Švihra J, Szépe P, Stehlik M (2021) "SPOCU'': scaled polynomial constant unit activation function. Neural Comput Appl 33(8):3385–3401
41. Ma D, Zhang F, Bull DR (2020) BVI-DVC: a training database for deep video compression. arXiv preprint arXiv:200313552
42. Moss FM, Wang K, Zhang F, Baddeley R, Bull DR (2015) On the optimal presentation duration for subjective video quality assessment. IEEE Trans Circuits Syst Video Technol 26(11):1977–1987
43. VTM 4.0 software. Available at: https://vcgit.hhi.fraunhofer.de/jvet/VVCSoftware_VTM/-/tree/VTM-4.0
44. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, Corrado GS, Davis A, Dean J, Devin M (2016) Tensorflow: large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:160304467
45. Da K (2014) A method for stochastic optimization. arXiv preprint arXiv:14126980
46. Ohm J-R, Sullivan GJ, Schwarz H, Tan TK, Wiegand T (2012) Comparison of the coding efficiency of video coding standards–including high efficiency video coding (HEVC). IEEE Trans Circuits Syst Video Technol 22(12):1669–1684
47. Bossen F, Boyce J, Li X, Seregin V, Sühring K (2019) JVET common test conditions and software reference configurations for SDR video. In: Document JVET-M1010, 13th JVET meeting, Marrakesh, pp 9–18
48. Bjontegaard G (2001) Calculation of average PSNR differences between RD-curves. VCEG-M33
49. Kawamura K, Kidani Y, Naito S (2019) CE13-2.6/CE13-2.7: evaluation results of CNN based in-Loop filtering. In: Document JVET-N0710, 14th JVET meeting, Geneva, Switzerland, pp 19–27
50. Wan S, Wang M, Ma Y, Huo J, Gong H, Zou C, et al (2019) CE10: integrated in-loop filter based on CNN (Tests 2.1, 2.2 and 2.3). In: The JVET meeting. Gothenburg, Sweden: ITU-T, ISO/IEC, number JVET-O0079
51. Huang Z, Li Y, Sun J (2020) Multi-gradient convolutional neural network based in-loop filter for VVC. In: 2020 IEEE international conference on multimedia and expo (ICME). IEEE, pp 1–6
52. DIV2K https://data.vision.ee.ethz.ch/cvl/DIV2K/