GAN-BASED EFFECTIVE BIT DEPTH ADAPTATION FOR PERCEPTUAL VIDEO COMPRESSION

Di Ma, Fan Zhang, David R. Bull

Department of Electrical and Electronic Engineering, University of Bristol, BS8 1UB, UK
{di.ma, fan.zhang, dave.bull}@bristol.ac.uk

ABSTRACT

Resolution and effective bit depth (EBD) adaptation have been recently utilised in video compression to improve coding efficiency. This type of approach dynamically reduces spatial/temporal resolutions and effective bit depth at the encoder and restores the original video formats during decoding. In this paper, a convolutional neural networks (CNN) based EBD adaptation method is presented for perceptual video compression, in which the employed CNN models are trained using a generative adversarial network (GAN), with perception-based loss functions. This method was integrated into the HEVC HM 16.20 reference software and fully evaluated on test sequences from the JVET Common Test Conditions using the Random Access configuration. The results show significant coding gains achieved on all test sequences with an overall bit rate saving of 24.8% (Bjøntegaard Delta measurement) based on a perceptual quality metric, VMAF.

Index Terms— Effective bit depth adaptation, CNN, GAN, HEVC, Perceptual video compression.

1. INTRODUCTION

In recent years the demand for more immersive video content with extended parameter spaces (higher spatial resolutions, frame rates, dynamic ranges and colour gamuts) has challenged the operational limits of video acquisition, compression and delivery [1]. In this context, ISO/IEC MPEG and ITU-T VCEG have jointly initialised a new video coding standard, Versatile Video Coding (VVC) [2], targeting 30-50% coding gains over the current standard, High Efficiency Video Coding (HEVC) [3]. Moreover, an open-source and royalty free video coding format, AOMedia Video 1 (AV1) [4] has been proposed by the Alliance for Open Media (AOM) [5] primarily for Internet streaming, which targets enhanced performance over and above HEVC [6].

To further improve coding efficiency, several researchers have exploited deep learning techniques - for intra [7, 8] and inter predictions [9,10], transforms [11,12], entropy coding [13,14], in-loop filtering [15,16] and format adaptation [17,18]. However the computational complexity for this type of approach can be high compared to conventional coding methods.

CNN-based format adaptation approaches adaptively reduce the spatial/temporal resolutions and effective bit depth during encoding, and apply up-sampling to reconstruct original video formats (resolutions and bit depth) at the decoder [19, 20]. Most of these methods [17, 19–22] have been trained to minimise pixel-wise distortions by using $\ell 1$ or $\ell 2$ lost functions. These however do not, in general, correlate well with subjective quality opinions [23].

Based on the recent work of using generative adversarial networks (GANs) for super-resolution processing [24, 25] and their application to video coding with spatial resolution adaptation [26], we propose an effective bit depth (EBD) adaptation approach using a perceptually trained CNN (BDGAN) for video compression. This employs residual dense blocks and cascading structures to effectively extract hierarchical features and maximise information flow. This method has been integrated into the HEVC HM 16.20 reference model for evaluation[1], achieving significant coding gains based on perceptual quality assessment compared to the original HM 16.20 on all JVET Common Test Conditions (CTC) test sequences with the Random Access (RA) configuration.

The remainder of the paper is organised as follows. The proposed effective bit depth adaptation approach is presented in Section 2 including a detailed description on CNN architecture and training strategy. Section 3 provides the compression results and complexity analysis with discussions. Finally, conclusions and future work are outlined in Section 4.

*The authors acknowledge funding from EPSRC (EP/L016656/1 and EP/M000885/1) and the NVIDIA GPU Seeding Grants.

[1]The proposed GAN-based EBD adaptation approach has not been tested on VVC test model (VTM), which was still under development at the time of writing.

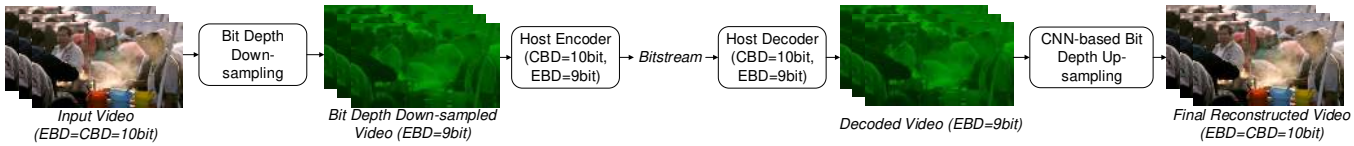


Fig. 1: Diagram of the proposed BDGAN-based Bit Depth Adaptation method.

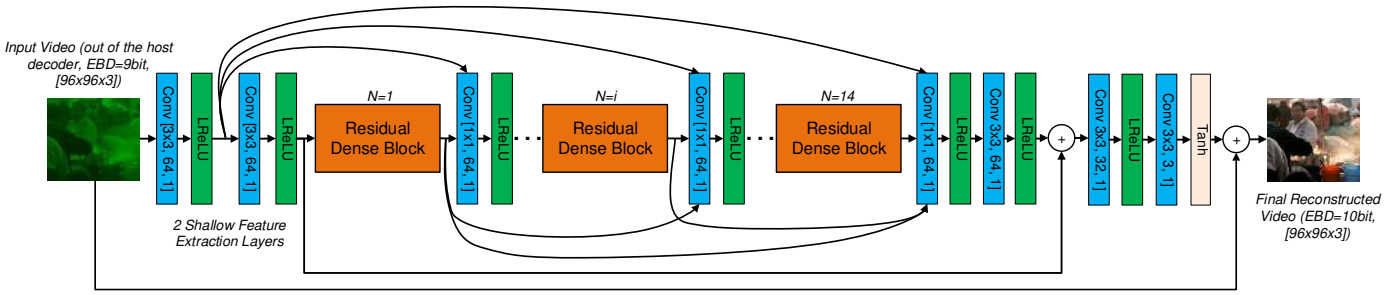


Fig. 2: Network architecture of the BDGAN's Generator (BDNet).

2. PROPOSED ALGORITHM

The coding framework with EBD adaptation is illustrated in Fig. 1, which is the same as reported in [20]. Here Effective Bit Depth (EBD) is the actual number of bits used to represent the video signal, which is different to Coding Bit Depth (CBD), which is defined as *InternalBitDepth* in HEVC HM and VVC VTM software.

Before encoding, the EBD of the original video frames is down-sampled by 1 bit through bit-shifting. The host encoder then compresses the video frames with reduced EBD (the CBD remains the same) to produce the bitstream. When receiving the bitstream, the host decoder reconstructs the reduced EBD video frames and applies the CNN-based up-sampling to obtain the final reconstructed frames with full EBD.

2.1. Quantisation Parameter Adjustment

In order to achieve meaningful comparison with anchor codecs without EBD adaptation, a quantisation parameter (QP) offset is applied on initial base QP values in this coding framework, which can produce similar bitrates with full EBD compression. This offset is fixed as -6 based on empirical observation and has been previously employed in similar applications [20, 27].

2.2. The Architecture of BDGAN

The proposed generative adversarial network architecture for EBD up-sampling (BDGAN) contains a generator (denoted as BDNet) and a discriminator. The BDNet structure is shown in Fig. 2, which takes a 96×96 YCbCr (4:4:4) decoded colour image block with EBD of 9 bit as input, and produces a 10 bit EBD image block in the same format, targeting to the corresponding original block.

The network starts with two shallow feature extraction layers, each of which contains a convolutional layer and a leaky ReLU (LReLU) [25] activation function. These are followed by 14 identical residual dense blocks (RDB) [28] as illustrated in Fig. 3. This RDB is different from the residual block used in SRGAN [24] - each convolutional layer in the RDB receives the feature maps from all preceding layers and feeds its output to subsequent layers. This structure preserves the feed-forward nature of the network and enables access to more previously extracted features, which improves the information flow [29]. An additional skip connection is employed in each RDB between the input of this RDB and the output of the last convolutional layer to further improve information flow and stabilise the training process [30].

Moreover, the output of the first shallow feature extraction layer and the output of each RDB are cascaded [31, 32] into all the following RDBs through a single 1×1 convolutional layer. After 14 RDBs, another convolutional layer is employed. A skip connection is applied between the output of this layer and the output of the second shallow feature extraction layer. Finally, another two convolutional layers are used before the final output, and an additional skip connection is employed between the input of these two layers and the initial BDNet input. The stride value for all convolutional layers in BDNet is 1.

Fig. 4 illustrates the BDGAN discriminator, which is identical to the SRGAN [24]. It has one shallow feature extraction

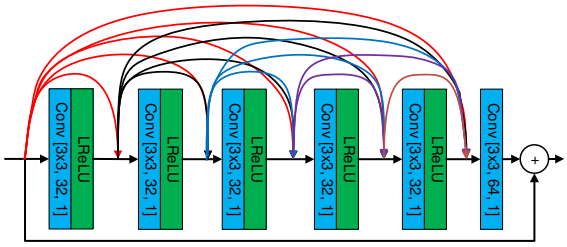


Fig. 3: Residual Dense Block (RDB) used in BDNet.

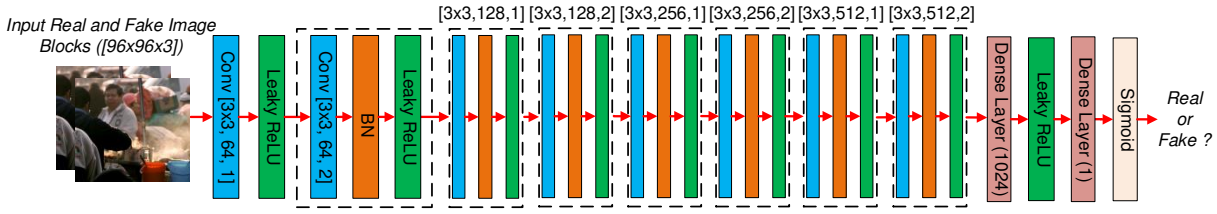


Fig. 4: Network architecture of the BDGAN's Discriminator.

layer, which is followed by seven convolutional layers with increased feature map numbers from 64 to 512. Two dense layers are then employed to produce the final binary output. The kernel size of each convolutional layer in the discriminator is 3×3 with a stride of 1 or 2.

2.3. Training Database

To effectively train the proposed CNN model and improve its generalisation, 432 source sequences at different spatial resolutions (2160p, 1080p, 540p and 270p) are employed, selected from publicly available databases [33–36]. All these original sequences are in 10bit YCbCr 4:2:0 format, and are segmented to 64 frames short clips without scene cut. The EBD of these sequences are then down-sampled to 9 bit through bit-shifting and compressed using the HEVC HM 16.20 software with the RA configuration of JVET CTC for initial base QP values, 22, 27, 32 and 37. Here the QP offset of -6 is applied on the initial base QPs as mentioned in Section 2.1.

2.4. Loss Functions

The training process is similar to that in [26], in which the BDGAN generator (BDNet) was firstly trained using the multi-scale structural similarity index (MS-SSIM) quality metric [37] as the loss function. Based on the optimal BDNet models obtained, the generator was trained again, together with the discriminator in order to generate more high frequency details [24] in the output. The loss function used here for training the generator, L_G, is given as below which combines the $\ell 1$ loss (MAD), SSIM loss [38] (L_{SSIM}), and the adversarial loss of the generator (L_G^{Ra}):

$$L_G = 0.025 \times \mathrm{MAD} + L_{\mathrm{SSIM}} + 5 \times 10^{-3} \times L_G^{Ra}. \tag{1}$$

where L_G^{Ra} is calculated by (2), based on the training strategy of Relativistic GANs [39].

$$\begin{aligned} L_G^{Ra} = &- E_{x_r}[\ln(1 - (\mathrm{Sig}(C_d(x_r) - E_{x_f}[C_d(x_f)])))] \\ &- E_{x_f}[\ln(\mathrm{Sig}(C_d(x_f) - E_{x_r}[C_d(x_r)]))] \end{aligned} \tag{2}$$

in which $E[\cdot]$ represents the mean operation, x_r and x_f are the real and fake image block respectively, and $C_d(\cdot)$ is the output of the discriminator of BDGAN. 'Sig' stands for the Sigmoid function.

For the discriminator, the loss function L_D is given by (3):

$$\begin{aligned} L_D = &- E_{x_r}[\ln(\mathrm{Sig}(C_d(x_r) - E_{x_f}[C_d(x_f)]))] \\ &- E_{x_f}[\ln(1 - (\mathrm{Sig}(C_d(x_f) - E_{x_r}[C_d(x_r)])))] \end{aligned} \tag{3}$$

2.5. Training Configuration

The compressed and original video frames were randomly selected and cropped to 96×96 image blocks in the YCbCr 4:4:4 format. In this process, block rotation is also employed to further increase the data diversity. This results in more than 90,000 pairs of blocks for each QP group.

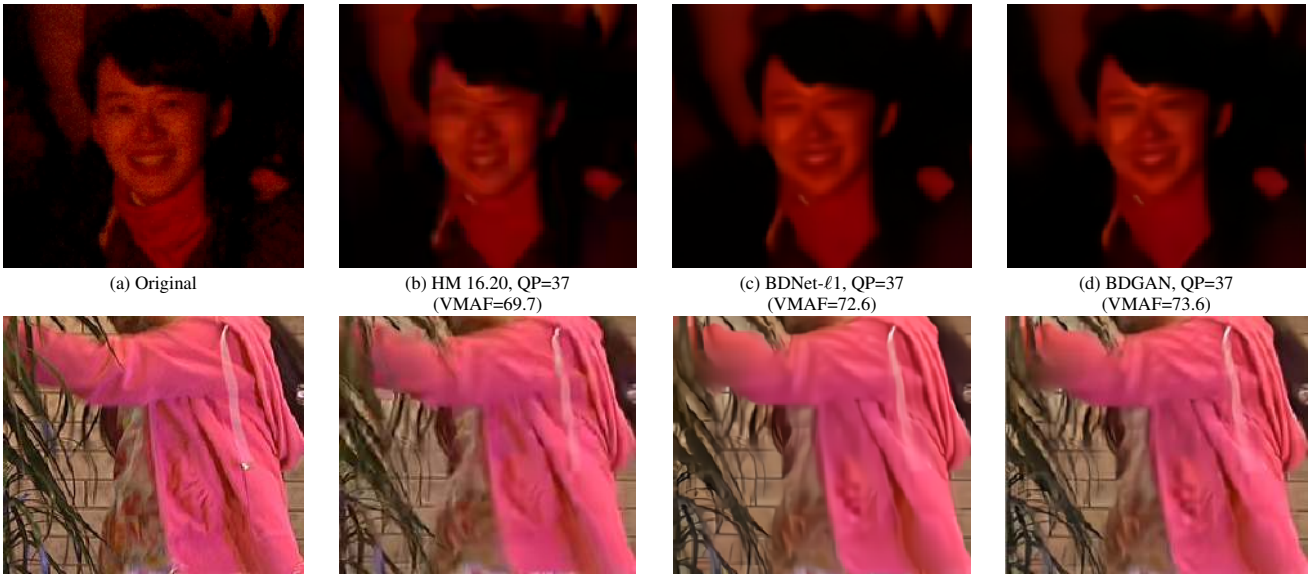
The CNN model was implemented and trained using the TensorFlow (version 1.8.0) based on the following parameters: Adam optimisation [40] with the hyper-parameters of β_1=0.9 and β_2=0.999; batch size of 16; 200 training epochs; learning rate (0.0001); weight decay of 0.1 for every 100 epochs.

For four initial base QP groups (22, 27, 32 and 37), the trained BDGAN models are denoted as below:

$$\text{BDGAN\_Models} = \begin{cases} \text{Model}_1, & \text{QP} \leq 24.5 \\ \text{Model}_2, & 24.5 < \text{QP} \leq 27.5 \\ \text{Model}_3, & 27.5 < \text{QP} \leq 32.5 \\ \text{Model}_4, & \text{QP} > 32.5 \end{cases} \tag{4}$$

3. RESULTS AND DISCUSSION

The proposed EBD adaptation approach has been integrated into the HEVC HM 16.20 reference software and evaluated on the JVET CTC test sequences using the Random Access configuration (Main10 profile) with initial base QP values 22, 27, 32 and 37. The test sequences used are taken from A1, A2, B, C and D video classes in JVET CTC, with spatial resolutions ranging from 2160p to 240p [41].



(a) Original (b) HM 16.20, QP=37 (VMAF=69.7) (c) BDNet-ℓ1, QP=37 (VMAF=72.6) (d) BDGAN, QP=37 (VMAF=73.6)

(e) Original (f) HM 16.20, QP=37 (VMAF=71.4) (g) BDNet-ℓ1, QP=37 (VMAF=74.4) (h) BDGAN, QP=37 (VMAF=76.9)

Fig. 5: Example blocks of the reconstructed frames for the anchor HM 16.20, EBD up-sampling with BDNet-ℓ1 and BDGAN (their bitstreams have similar bit rates). These are from the 175th and 162nd frames of *Campfire* and *PartyScene* sequences respectively and amplified by 4 times.

The proposed approach was compared to the original HEVC HM 16.20 using the Bjøntegaard Delta [42] measurement (BD-rate). Here video quality was assessed by two quality metrics, PSNR (Peak Signal to Noise Ratio-applying on luminance channel only) and VMAF (Video Multi-Method Assessment Fusion-version 0.6.1) [43]. PSNR is the most widely used quality metric for compression, although it does not correlate very well with subjective quality [44]. VMAF employs a machine learning approach combining multiple quality metrics and video feature together, which improves its correlation with visual quality [23].

The computational complexity of this work was estimated by calculating the average encoding and decoding time. Both the encoder and the decoder are executed on a PC with an Intel(R) Core(TM) i7-4770K CPU @3.5GHz and a NVIDIA P6000 GPU device (24GB RAM).

3.1. Compression Performance

Table 1 summarises the compression performance of the proposed method, with HM 16.20 used as benchmark. In order to highlight the improvement obtained by using perceptual loss functions, the coding gains from an $\ell 1$ trained (using the same training material) BDNet for EBD up-sampling (BDNet-$\ell 1$) are also reported for comparison.

It is noted that EBD adaptation with BDGAN achieves consistent coding gains for all test sequences, with average BD-rate (assessed by VMAF) of 24.8%, which is 7.4% higher than that for BDNet-$\ell 1$. When PSNR is employed for video quality assessment, the savings are much lower and not as significant as those for BDNet-$\ell 1$. Comparing to [20], where only PSNR-based results were presented, additional 3.9% coding gains have been achieved by using BDNet-$\ell 1$. This is due to the more advanced CNN model employed and the large video database used for training.

The bitrate-quality (PSNR and VMAF) curves of the anchor HM 16.20 and the proposed method (using BDNet-$\ell 1$ and BDGAN for bit depth up-sampling) for selected sequences, *Campfire* (Class A), *Cactus* (Chass B), *PartyScene* (Chass C) and *BQSquare* (Class D), are shown in Fig. 6.

Table 1: BD-rate results for JVET CTC tested sequences.

Sequence	BDNet-$\ell 1$		BDGAN	
	BD-Rate (PSNR)	BD-Rate (VMAF)	BD-Rate (PSNR)	BD-Rate (VMAF)
A1-Campfire	-16.7%	-25.3%	-14.8%	-29.7%
A1-FoodMarket4	-7.0%	-13.5%	-3.2%	-14.7%
A1-Tango2	-9.2%	-16.9%	-5.3%	-19.2%
A2-CatRobot1	-12.0%	-22.2%	-7.4%	-29.6%
A2-DaylightRoad2	-14.4%	-25.0%	-7.8%	-35.7%
A2-ParkRunning3	-14.2%	-19.9%	-11.9%	-28.0%
Class A (2160p)	**-12.3%**	-20.5%	-8.4%	**-26.2%**
B-BasketballDrive	-10.7%	-14.4%	-6.3%	-20.6%
B-BQTerrace	-10.2%	-28.8%	-5.3%	-41.1%
B-Cactus	-10.2%	-18.7%	-5.9%	-30.0%
B-MarketPlace	-4.6%	-13.7%	-1.4%	-25.3%
B-RitualDance	-7.1%	-13.3%	-3.5%	-19.4%
Class B (1080p)	**-8.6%**	-11.0%	-4.5%	**-27.3%**
C-BasketballDrill	-9.2%	-14.6%	-5.3%	-21.7%
C-BQMall	-9.0%	-13.7%	-4.9%	-22.3%
C-PartyScene	-7.7%	-12.5%	-2.9%	-22.9%
C-RaceHorses	-8.7%	-14.1%	-5.0%	-22.0%
Class C (480p)	**-8.7%**	-13.7%	-4.5%	**-22.2%**
D-BasketballPass	-10.3%	-12.7%	-7.0%	-18.7%
D-BlowingBubbles	-7.5%	-12.6%	-3.7%	-22.3%
D-BQSquare	-17.0%	-23.9%	-9.8%	-26.1%
D-RaceHorses	-9.7%	-14.4%	-6.3%	-22.0%
Class D (240p)	**-11.1%**	-15.9%	-6.7%	**-22.3%**
Overall	**-10.3%**	-17.4%	-6.2%	**-24.8%**

The example blocks from the reconstructed frames generated by the anchor HM 16.20, and EBD up-sampling with BDNet-$\ell 1$ and BDGAN are shown in Fig. 5. It can be observed that for both BDGAN and BDNet-$\ell 1$, the reconstructed content exhibits improved perceptual quality, with fewer blocking artefacts compared to the anchor. BDGAN results also exhibit more texture detail and higher contrast than those for BDNet-$\ell 1$.

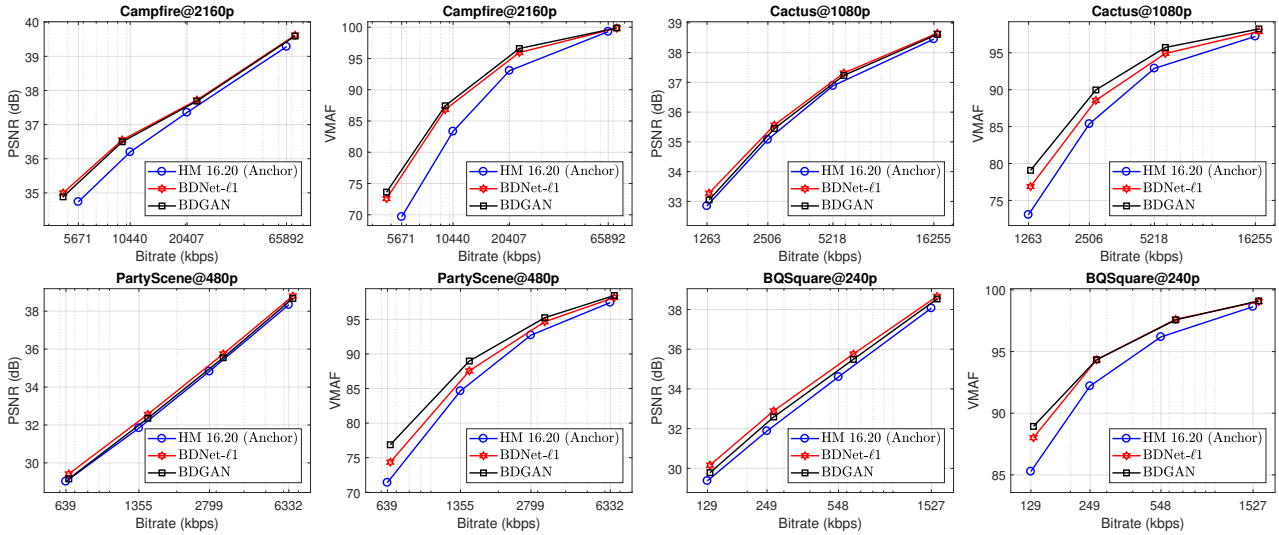


Fig. 6: Rate-quality curves of the anchor, BDNet-$\ell 1$ and BDGAN networks.

3.2. Complexity Analysis

The average encoding time of the proposed approach is 1.015 times that of the anchor HM 16.20, while the average decoding time (the same for BDNet-$\ell 1$ and BDGAN) is 87.4 times that of the original HM decoder, which is higher than that in [20]. This is due to the more complex CNN model used for bit depth up-sampling.

4. CONCLUSION

In this paper, an effective bit depth (EBD) adaptation approach has been presented for perceptual video coding. The EBD up-sampling is implemented using a GAN-based model (BDGAN), which was trained with perceptually-inspired loss functions. This method has been integrated with HEVC HM 16.20 and fully tested on JVET CTC test sequences. The compression results show significant coding gains achieved for all test sequences, assessed using the perceptual quality metric VMAF, with an average BD-rate value of 24.8%. Future work will focus on complexity reduction for the CNN and the integration with the emerging VVC standard.

5. REFERENCES

[1] D. R. Bull, *Communicating pictures: A course in Image and Video Coding*, Academic Press, 2014.

[2] B. Bross, J. Chen, S. Liu, and Y.-K. Wang, "Versatile video coding (draft 7)," in *JVET-P2001. ITU-T and ISO/IEC*, 2019.

[3] ITU-T Rec. H.265, "High efficiency video coding," ITU-T Std., (2015).

[4] AOM, "AOMedia Video 1 (AV1)," https://aomedia.googlesource.com/.

[5] "Alliance for Open Media," https://aomedia.org/.

[6] A. V. Katsenou, F. Zhang, M. Afonso, and D. R. Bull, "A subjective comparison of AV1 and HEVC for adaptive video streaming," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 4145–4149.

[7] W. Cui, T. Zhang, S. Zhang, F. Jiang, W. Zuo, and D. Zhao, "Convolutional neural networks based intra prediction for HEVC," *arXiv preprint arXiv:1808.05734*, 2018.

[8] S. Kuanar, K. Rao, and C. Conly, "Fast mode decision in HEVC intra prediction, using region wise CNN feature classification," in *2018 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, 2018, pp. 1–4.

[9] H. Zhang, L. Li, L. Song, X. Yang, and Z. Li, "Advanced cnn based motion compensation fractional interpolation," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 709–713.

[10] Y. Wang, X. Fan, C. Jia, D. Zhao, and W. Gao, "Neural network based inter prediction for HEVC," in *2018 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2018, pp. 1–6.

[11] S. Puri, S. Lasserre, and P. Le Callet, "CNN-based transform index prediction in multiple transforms framework to assist entropy coding," in *2017 25th European Signal Processing Conference (EUSIPCO)*. IEEE, 2017, pp. 798–802.

[12] S. Jimbo, J. Wang, and Y. Yashima, "Deep learning-based transformation matrix estimation for bidirectional interframe prediction," in *2018 IEEE 7th Global Conference on Consumer Electronics (GCCE)*. IEEE, 2018, pp. 726–730.

[13] R. Song, D. Liu, H. Li, and F. Wu, "Neural network-based arithmetic coding of intra prediction modes in HEVC," in *2017 IEEE Visual Communications and Image Processing (VCIP)*. IEEE, 2017, pp. 1–4.

[14] C. Ma, D. Liu, X. Peng, and F. Wu, "Convolutional neural network-based arithmetic coding of DC coefficients for HEVC intra coding," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 1772–1776.

[15] Y. Wang, H. Zhu, Y. Li, Z. Chen, and S. Liu, "Dense residual convolutional neural network based in-loop filter for HEVC," in *2018 IEEE Visual Communications and Image Processing (VCIP)*. IEEE, 2018, pp. 1–4.

[16] T. Li, M. Xu, C. Zhu, R. Yang, Z. Wang, and Z. Guan, "A deep learning approach for multi-frame in-loop filter of HEVC," *IEEE Transactions on Image Processing*, vol. 28, no. 11, pp. 5663–5678, 2019.

[17] Y. Li, D. Liu, H. Li, L. Li, F. Wu, H. Zhang, and H. Yang, "Convolutional neural network-based block up-sampling for intra frame coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 9, pp. 2316–2330, 2018.

[18] F. Zhang, M. Afonso, and D. R. Bull, "Vistra2: Video coding using spatial resolution and effective bit depth adaptation," *arXiv preprint arXiv:1911.02833*, 2019.

[19] M. Afonso, F. Zhang, and D. R. Bull, "Video compression based on spatio-temporal resolution adaptation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 1, pp. 275–280, 2019.

[20] F. Zhang, M. Afonso, and D. R. Bull, "Enhanced video compression based on effective bit depth adaptation," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019.

[21] J. Lin, D. Liu, H. Yang, H. Li, and F. Wu, "Convolutional neural network-based block up-sampling for HEVC," *IEEE Transactions on Circuits and Systems for Video Technology*, 2018.

[22] M. Afonso, F. Zhang, and D. R. Bull, "Spatial resolution adaptation framework for video compression," in *Applications of Digital Image Processing XLI*. International Society for Optics and Photonics, 2018, vol. 10752, p. 107520L.

[23] F. Zhang, F. M. Moss, R. Baddeley, and D. R. Bull, "BVI-HD: A video quality database for HEVC compressed and texture synthesized content," *IEEE Transactions on Multimedia*, vol. 20, no. 10, pp. 2620–2630, 2018.

[24] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al., "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4681–4690.

[25] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. Change Loy, "ESRGAN: Enhanced super-resolution generative adversarial networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 0–0.

[26] D. Ma, M. F. Afonso, F. Zhang, and D. R. Bull, "Perceptually-inspired super-resolution of compressed videos," in *Applications of Digital Image Processing XLII*. International Society for Optics and Photonics, 2019, vol. 11137, pp. 310–318.

[27] M. Afonso, F. Zhang, A. Katsenou, D. Agrafiotis, and D. Bull, "Low complexity video coding based on spatial resolution adaptation," in *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2017, pp. 3011–3015.

[28] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2472–2481.

[29] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[31] N. Ahn, B. Kang, and K.-A. Sohn, "Fast, accurate, and lightweight super-resolution with cascading residual network," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 252–268.

[32] N. Ahn, B. Kang, and K.-A. Sohn, "Photo-realistic image super-resolution with fast and lightweight cascading residual network," *arXiv preprint arXiv:1903.02240*, 2019.

[33] A. Mackin, F. Zhang, and D. R. Bull, "A study of subjective video quality at various frame rates," in *2015 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2015, pp. 3407–3411.

[34] M. A. Papadopoulos, F. Zhang, D. Agrafiotis, and D. Bull, "A video texture database for perceptual compression and quality assessment," in *2015 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2015, pp. 2781–2785.

[35] I. Katsavounidis, "Chimera video sequence details and scenes," tech. rep., Netflix, (November 2015).

[36] Harmonic Inc 4K demo footage, ," http://www.harmonicinc.com/4k-demo-footage-download/, (Accessed: 1st May 2017).

[37] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*. Ieee, 2003, vol. 2, pp. 1398–1402.

[38] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, et al., "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[39] A. Jolicoeur-Martineau, "The relativistic discriminator: a key element missing from standard GAN," *arXiv preprint arXiv:1807.00734*, 2018.

[40] D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[41] F. Bossen, J. Boyce, X. Li, et al., "JVET common test conditions and software reference configurations for sdr video," *Joint Video Experts Team (JVET) of ITU-T SG*, vol. 16, 2018.

[42] G. Bjøntegaard, "Calculation of average PSNR differences between rd-curves," in *13th VCEG Meeting, no. VCEG-M33,Austin, Texas*, 2001, pp. USA: ITU–T.

[43] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Manohara, "Toward a practical perceptual video quality metric," *The Netflix Tech Blog*, vol. 6, 2016.

[44] F. Zhang and D. R. Bull, "A perception-based hybrid model for video quality assessment," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 6, pp. 1017–1028, 2015.