The International Journal of Engineering and Science (IJES)
|| Volume || 10 || Issue || 10 || Series I || Pages || PP 01-07 || 2021 ||
ISSN (e): 2319-1813 ISSN (p): 20-24-1805

A Deep Learning Approach for HEVC Intra Coding

Tien-Yang Hsu, Tung-Hung Hsieh, Yueh-Ju Lu, Chou-Chen Wang
Department of Electronic Engineering, I-Shou University, Kaohsiung, Taiwan
Corresponding Author: Chou-Chen Wang

----------ABSTRACT----------
Convolutional neural network (CNN) has been developed rapidly in deep learning areas, and has become a hot studying topic in image applications. High efficiency video coding (HEVC) is the newest video coding standard. The HEVC standard can achieve high coding efficiency with a lower bitrate for intra frame coding. However, it still needs many bits to finish best rate-distortion (R-D) curve. Since there are only 35 directions prediction modes provided in intra prediction module (IPM), HEVC occurs a large distortion when the image contents are out of these prediction directions. In other words, HEVC can achieve high coding performance when the image contents match these 35 prediction directions. In order to obtain a better R-D curve, Zhang et al. [3] recently proposed a simple CNN (S-CNN) to improve the encoding performance of HEVC. The S-CNN consists of super-resolution CNN (SRCNN) [4] and ResNet [5] with two layer networks. The S-CNN can precisely predict the residual information of coding tree unit (CTU) and achieve a better R-D performance for HEVC encoder. However, S-CNN has to consume more time to encode intra frame coding since it needs to perform more CNN enhancement mode. In order to further speed up S-CNN based intra frame coding, we propose an early termination algorithm to skip CNN. Because the natural images have high spatial correlation, we find the mean square errors (MSE) of reconstructed CTU exist high spatial correlation at HEVC encoder. Therefore, a dynamic threshold of MSE is set according to four neighboring encoded CTU blocks to evaluate whether the current reconstructed CTU is useful for the CNN enhancement mode. Simulation results show that both our proposed method and S-CNN can reach better R-D curves. On the other hand, although our proposed method increases 0.9% and loses 0.05 dB in average BD-BitRate and BD-PSNR as compared with S-CNN, respectively. However, we can achieve faster HEVC encoding process than S-CNN by reducing time increase ratio (TIR) about 13% on an average.
KEYWORDS– *High efficiency video coding, deep learning, convolutional neural network.*

Date of Submission: 29-09-2021 Date of Acceptance: 12-10-2021

I. INTRODUCTION

The high efficiency video coding (HEVC) is a very popular video encoding standard for 4K/8K ultrahigh definition (UHD) video applications [1-2]. This is because HEVC standard can provide better video quality with a lower bitrate [2]. Therefore, the higher coding efficiency of HEVC can make it more suitable for UHD video. HEVC mainly consists of intra frame coding and inter frame coding based on spatial and temporal correlations. In HEVC encoder, the intra frame coding plays a very important role to obtain optimal rate-distortion (R-D) curve. The intra frame coding provides a flexible quadtree structure, which varies from 64×64 to 8×8 pixels, to achieve the best intra frame predicting coding to reconstruct frame. Therefore, how to increase the efficiency of intra frame coding in HEVC is a popular studying issue.

Recently, a deep convolutional neural network (CNN) has achieved a great success in multimedia information processing fields and pattern recognition technology. Therefore, many studies [3-5] has been proposed to improve the performance of intra frame coding using a CNN-based method. In order to obtain a better R-D curve, Zhang et al. [3] recently proposed a simple CNN (S-CNN) to improve the encoding performance of HEVC. The S-CNN consists of super-resolution CNN (SRCNN) [4] and ResNet [5] with two layer networks. The S-CNN can precisely predict the residual information of coding tree unit (CTU) and achieve a better R-D performance for HEVC encoder. However, S-CNN has to consume more time to encode intra frame coding since it needs to perform more CNN enhancement mode.

In order to further speed up S-CNN based intra frame coding, we propose an early termination algorithm to skip CNN. Because the natural images have stationary characteristics, we find the mean square errors (MSE) of reconstructed CTU exist high spatial correlation in intra frame. Therefore, a dynamic threshold of MSE is set according to four neighboring encoded CTU blocks to evaluate whether the current reconstructed CTU is useful for the CNN enhancement mode.

The remainder of this paper is organized as follows. In Section II we briefly review related approaches for CNN-based HEVC intra frame coding. Section III elaborates the proposed early termination algorithm to skip CNN. The experimental results are presented in Section IV. Finally, Section V summarizes our conclusions.

II. BRIEF REVIEWS OF HEVC AND CNN-BASED INTRA FRAME CODING

2.1 HEVC Encoder

HEVC encoder mainly consists of three modules including coding unit (CU), prediction unit (PU) and transform unit (TU), as shown in Fig. 1 [1-2]. The CU is the basic unit of region splitting used for inter/intra prediction, which allows recursive subdividing into four equally sized blocks. The CU can be split by coding quadtree-structured CTU structure of 4 level depths, which CU size range from largest CU (64×64) to the smallest CU (8×8) pixels. At each depth level (CU size), HEVC performs motion estimation (ME) and motion compensation (MC) with different size. The PU is the basic unit used for carrying the information related to the prediction processes, and the TU can be split by residual quad-tree (RQT) at maximally 3 level depths which vary from 32×32 to 4×4 pixels. In general, intra-coded CUs have only two PU partition types including 2N×2N and N×N. The rate distortion costs (RDcost) have to be calculated by performing the PUs and TUs to select the optimal partition mode under all partition modes for each CU size. In intra frame coding, there are 35 angular intra prediction modes (IPM) based on symmetrical relationship adopted in HEVC standard. In other words, HEVC intra prediction processes various sizes of PUs and 35 prediction modes to generate a set of prediction pixels for each PU.

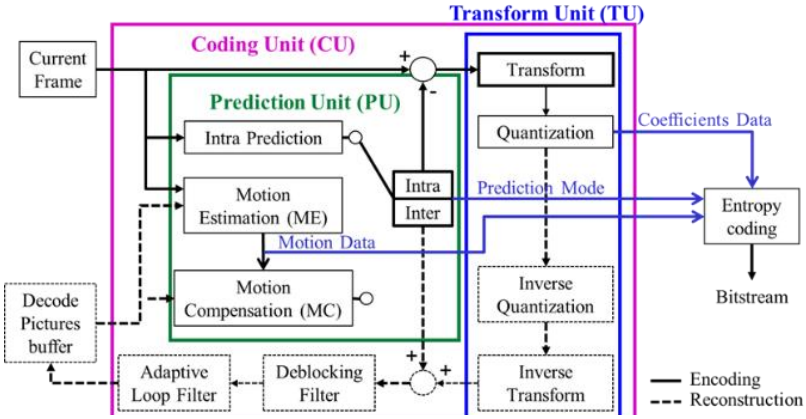
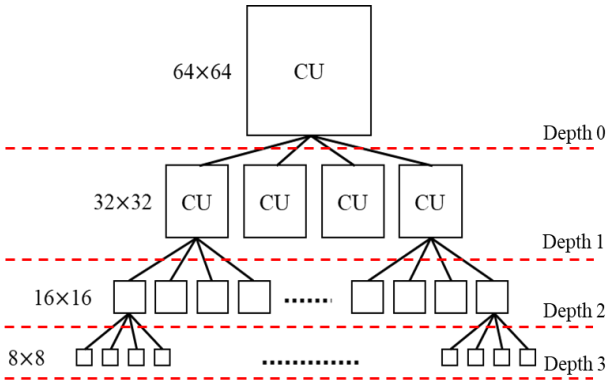


Fig.1 The block diagram of HEVC encoder.

Figure 2 shows the computational process for a CU to find the optimal partition mode from IPM module in HEVC. Although the IPM can enhance the PU performance and allow the encoder to search a better intra frame from several previous encoded CTUs, the computational complexity of the IPM dramatically increases. In Fig. 1, we also summarize the complexity of IPM module in HM 16.7 [6] test platform using 3,840×2,160 (4K) video sequence for per intra frame. From Fig. 1, we can find that the total number of IPM calculations required in one intra frame is up to 6,024,375 times. Therefore, high computational complexity becomes a bottleneck for the real-time applications of HEVC in UHD videos.



Depth	CU/CTU	Modes/CU	CTUs/frame	IPM times
0	1	35	2025	70,875
1	4	35	2025	283,500
2	16	35	2025	1,134,000
3	64	35	2025	4,536,000
Total IPM times				6,024,375

Fig.2 The computational complexity of IPM.

2.2 SRCNN and ResNet

The basic architecture of CNN is composed of one or more convolutional layers and pooling layers. For image applications, the CNN method can be regarded as an image-input and image-output structure. Therefore, Dong et al. [4] proposed a SRCNN with fully convolutional neural network for single image to super-resolution. They perform a low-resolution image using a bilinear interpolation to get a high-resolution image which is a ground truth image. And then, the SRCNN directly learns end-to-end mapping between the low and high-resolution images. SRCNN mainly consists of the three stages including extraction of feature maps, non-linear mapping, and image reconstruction. However, the non-linear mapping step would occupy the main computational complexity of the overall super-resolution process due to the high-dimensional vector mapping.

On the other hand, residual learning is a popular learning strategy in deep learning recently. Based on ResNet, a residual learning framework would alleviate the training of deep networks [5]. The main idea is that instead of hoping each stack of layers directly fits a desired underlying mapping, the residual network explicitly lets these layers fit approximate a residual function. In a deep ResNet framework, the output function is denoted as $F(Y)$ when the input Y. In general, denoting the desired underlying mapping as $H(Y)$, we let the stacked nonlinear layers fit another mapping of $F_{res}(Y) = H(Y) - Y$. Figure 3 shows the formulation of $F_{res}(Y) + Y$ can be realized by feedforward neural networks with shortcut connections.

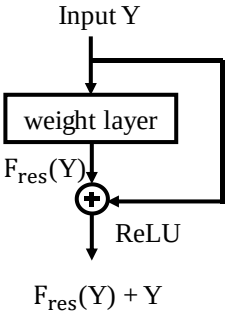


Fig.3 A building block of residual learning.

2.3. CNN-based intra frame coding

In order to obtain a better R-D curve of HEVC, Zhang et al. proposed S-CNN algorithm to improve the performance of intra frame coding. The S-CNN consists of SRCNN and ResNet with two layer networks, as shown in Fig.4. The S-CNN simplify the CNN-based structure for image reconstruction to only two fully convolutional layers which are feature map extraction and image reconstruction. In S-CNN based HEVC encoding of I-frame, each reconstructed CTU is generated after intra prediction and residual coding, and then the enhancement mode of S-CNN is used to reconstruct CTU. In this stage, the reconstructed CTU has enhanced visual quality from the predicted residual from the S-CNN mode. It is noted that the predicted residual from the CNN model is totally different from the residual form of intra prediction. And then, the new reconstructed CTU replaces the original one as the reference block for intra prediction of the next CTU to be encoded. Finally, a new reconstructed image can be obtained by the function of $H(Y) = F_{res}(Y) + Y$.

Therefore, the S-CNN can precisely predict the residual information of CTU and achieve a better R-D performance for HEVC encoder. However, S-CNN has to consume more time to encode intra frame coding since it needs to perform more CNN enhancement mode.

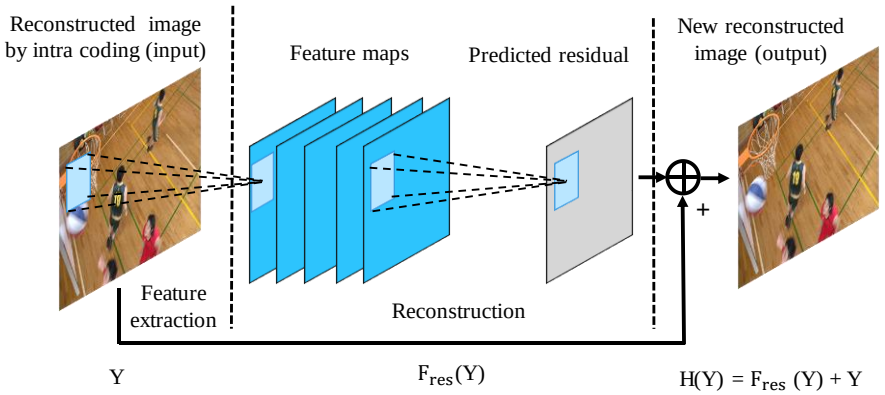


Fig.4 S-CNN architecture.

III. PROPOSED METHOD

The process of IPM module is shown in Fig. 5. Firstly, the I-frame is split into non-overlapping CTUs. Secondly, the IPM performs intra coding to get the prediction of CTU0 according to 35 prediction direction modes. And then, the reconstructed CTU_R replaces the CTU0. Finally, the residual information of CTU is encoded to bitstream. And so on, a reconstructed image is decoded. In intra frame coding, IPM provides 33 angular directions based on symmetrical relationship for prediction. Figure 5 shows the mapping relationship between the predicted mode and image pattern. From Fig.6, we can observe the predicted mode 17 fully matches the direction characteristics of line pattern in CTU image. It can be seen that if the image characteristics of direction meet the 35 prediction modes, HEVC encoder can finish a better reconstructed image. In other words, the better the reconstructed CTU obtained through IPM, the better the result of the next CTU prediction. On the other hand, the reconstructed CTU will be poor and affected when the image pattern is relatively inconsistent with 35 prediction modes, such as the direction of an arc in the image.

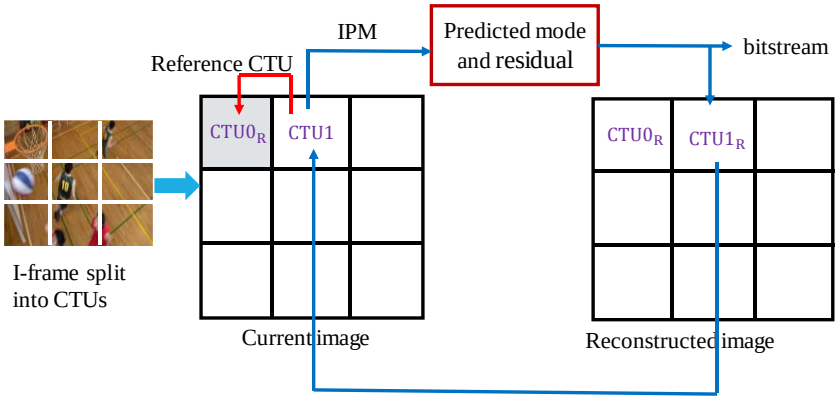


Fig.5 The process of IPM module.

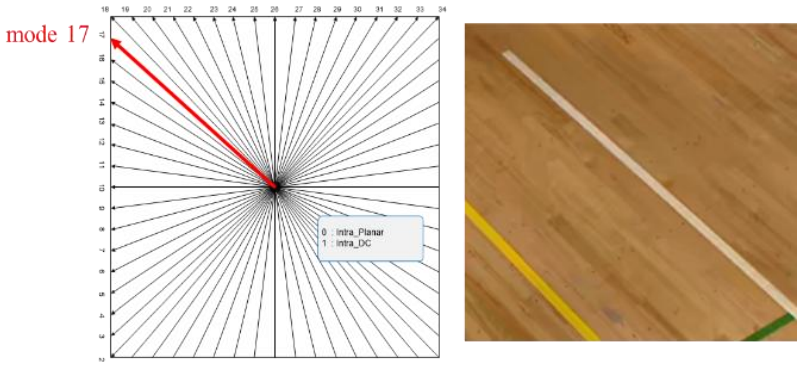


Fig.6 The mapping relationship.

In order to further speed up S-CNN based intra frame coding, we propose an early termination algorithm to skip CNN. In general, natural video sequences have strongly spatial and temporal correlations, especially in the homogeneous regions. Therefore, the MSE of reconstructed CTU is very close to the MSE of its spatially neighboring reconstructed CTUs due to the high correlation between adjacent CTUs, , as shown in Fig.7. Therefore, a dynamic threshold of MSE is set according to three neighboring encoded CTU blocks to evaluate whether the current reconstructed CTU is useful for the CNN enhancement mode.

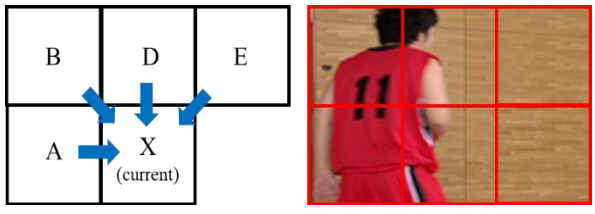


Fig.7 Four spatial neighboring reconstructed CTUs.

To employ the spatial correlation of MSE, we adopt the average of four spatial four neighboring CTUs in the current frame. Figure 8 shows the four neighboring MSE of encoded CTU for current CTU (X), including left (A_{MSE}), left upper (B_{MSE}), upper (D_{MSE}) and right upper (E_{MSE}), respectively. To decide whether early terminate the CNN, a threshold Thr_{MSE} of MSE is set to an average value defined as follows

$$Thr_{MSE} = \frac{A_{MSE}+B_{MSE}+D_{MSE}+E_{MSE}}{4} \quad (1)$$

As a result, the Thr_{MSE} is used to evaluate whether the current reconstructed CTU is helpful for the CNN enhancement mode in S-CNN. When the MSE is lower than the Thr_{MSE}, the CNN enhancement mode is skipped. The flowchart of the proposed early termination algorithm to skip CNN enhancement mode is shown in Fig. 9.

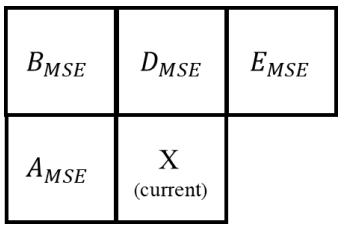


Fig.8 The neighboring MSE of current CTU.

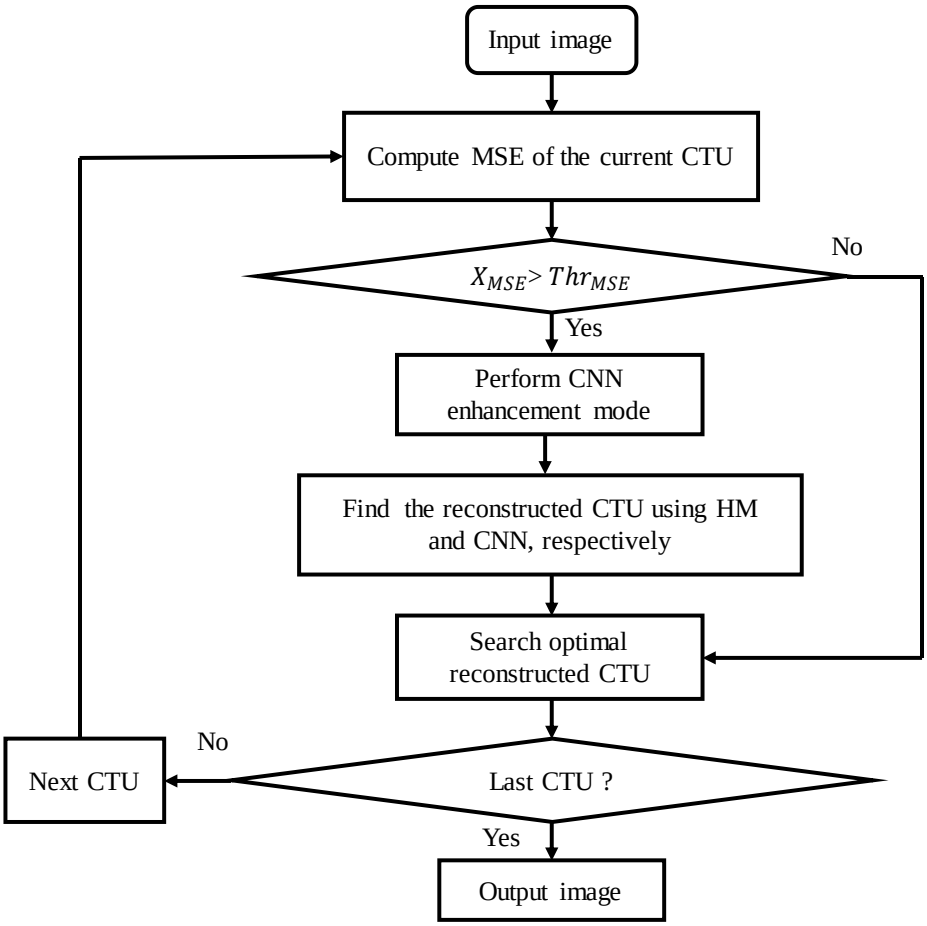


Fig.9 The flowchart of the proposed method.

IV. EXPERIMENTAL RESULTS

In this paper, we have implemented an early termination algorithm based on GPU in HM16.7 [6] encoder test model, the encoding configuration is summarized as follows:

(1) Scenario: all intra frames.
(2) QP = 27, 32, 37, 42
(3) Max partition depth: 3
(4) To be encoded frames: 10 frames
(5) Standard sequences:
Class A (2560×1600)
Class B (1920×1080)
Class C (832×480)
Class D (416×240)
Class E (1280×720)

Simulations are conducted on a desktop with

(1) OS: Windows 10 64-bit
(2) CPU: Intel Core i5-3470
(3) GPU: NVIDIA GeForce GTX 1060-3GB
(4) Memory: 12 GB
(5) Python version: Python3.6.5

In our experiments, the proposed CNN model is trained by four times corresponding to different QP values, respectively. For a fair comparison, we selected the training samples with four fixed ranges of MSE values in HEVC intra frame coding as the same conditions of Zhang's method [3]. Table I shows the different ranges of MSE values accordingly for extracting the corresponding training sample pairs. And, the training parameter settings are summarized in Table II. The coding performance is evaluated by the comparisons of BD-Bitrate (Bjontegaard delta bit rate) and BD-PSNRY (Bjontegaard delta peak signal-to-noise rate) [8]. On the other hand, to evaluate the time improvement of the proposed method, we define CNN-Ratio of performing CNN modes, IPM time increasing ratio (TIR_{IPM}) and HEVC encoding time increasing ratio (TIR_{HEVC}) as follows :

$$CNN\text{-}Ratio = \frac{N_{Proposed}}{N_{Zhang}} \times 100\% \quad (1)$$

$$TIR_{IPM} = \frac{IPM\text{-}CNN_{time}}{IPM\text{-}HM_{time}} \times 100\% \quad (2)$$

$$TIR_{HEVC} = \frac{Method_{time}}{HM_{time}} \times 100\% \quad (3)$$

TABLE I. The different ranges of MSE values for corresponding training sample.

QP	27	32	37	42
MSE	[1,5]	[5,15]	[15,50]	[50,200]

TABLE II. Parameter settings for CNN training.

input size	32×32
label size	32×32
data set	RAISE
training samples	150,000
filter size	5×5
active function	ReLU
Learning rate	$10^{-4} \sim 10^{-5}$

Table III and Table IV show the time increasing ratio and R-D performance comparisons between the proposed method and Zhang's method [3], respectively. As shown in the Tables III, the proposed method can skip about 53% CNN enhancement modes ratio when compared as Zhang's method. In other words, our method can perform IPM module faster than Zhang's method due to less time increasing ratio (TIR_{IPM}). Simulation results show that both our proposed method and Zhang's method can reach better R-D curves, as shown in Table IV. Although our proposed method increases 0.9% and loses 0.05 dB in average BD-BitRate and BD-PSNR as compared with Zhang's method, respectively. However, we can achieve faster HEVC encoding process than Zhang's S-CNN by reducing TIR_{HEVC} about 13% on an average.

TABLE III. Comparisons of time increasing ratio for IPM module.

Sequence		Average of CNN			TIR_{IPM}	
		N_{Zhang}	$N_{Proposed}$	CNN-Ratio	Zhang	Proposed
Class A	Traffic	10000	4580.25	45.8%	117	107
	PeopleOnStreet	10000	5195	51.95%	117	110
Class B	BasketballDrive	5100	2357.25	46.22%	117	107
	BQTerrace	5100	2726.75	53.47%	117	109
Class C	BasketballDrill	1040	541	52.04%	117	108
	PartyScene	1040	602	57.96%	117	110
Class D	BasketballPass	280	149.5	53.39%	117	109
	BQSquare	280	154.75	55.27%	117	109
Class E	FourPeople	2400	1184.5	49.35%	117	108
	KristenAndSara	2400	1173.75	48.91%	117	108
Average				52.33%	117	108

TABLE IV. Comparisons of R-D performance.

Sequence		BD-Bitrate(%)		BD-PSNR(dB)		TIR_{HEVC}(%)	
		Zhang	Proposed	Zhang	Proposed	Zhang	Proposed
Class A	Traffic	-3.6	-2.1	0.17	0.1	128	113
	PeopleOnStreet	-3.3	-2	0.17	0.11	130	116
Class B	Kimono1	-1.8	-1.1	0.08	0.05	131	116
	ParkScene	-2.3	-1.4	0.08	0.04	127	113
	Cactus	-1.9	-1.1	0.07	0.04	125	112
	BasketballDrive	-3	-2.1	0.1	0.06	128	113
	BQTerrace	-2	-1.4	0.1	0.07	126	114
Class C	BasketballDrill	-3.5	-2.3	0.13	0.09	122	111
	BQMall	-2.1	-1.4	0.11	0.08	120	112
	PartyScene	-1.6	-1	0.08	0.06	117	110
	RaceHorses	-2.4	-1.3	0.12	0.07	120	109
Class D	BasketballPass	-2.3	-1.7	0.11	0.09	128	115
	BQSquare	-1.3	-0.7	0.07	0.03	122	112
	BlowingBubbles	-1.7	-1	0.08	0.04	124	116
	RaceHorses	-3.2	-1.2	0.13	0.03	125	113
Class E	FourPeople	-4	-2.8	0.21	0.15	128	114
	Johnny	-3.6	-2.4	0.14	0.08	130	117
	KristenAndSara	-3.2	-2	0.16	0.09	132	116
Average		-2.6	-1.61	0.12	0.07	126	113

V. CONCLUSION

In this paper, we presented an efficient CNN-based intra frame coding for HEVC encoder. Our results indicate that the proposed method outperforms Zhang's S-CNN method in time increasing ratio and R-D performance.

REFERENCES

[1]. G. J. Sullivan, J-R Ohm, W-J Han, T. Wiegand, "Overview of the High Efficiency Video Coding (HEVC) Standard". *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 22, pp. 1649-1668, Dec. 2012.

[2]. High Efficiency Video Coding, Rec. ITU-T H.265 and ISO/IEC 23008-2, Jan. 2013.

[3]. Z. T. Zhang, C. H. Yeh, L. W. Kang, and M. H. Lin, "Efficient CTU-based intra frame coding for HEVC based on deep learning," *Proc. Asia Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, pp. 661-664, Dec. 2017.

[4]. C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pp. 184-199, 2014.

[5]. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 770-778, Jun. 2016.

[6]. HEVC test platform. https://hevc. hhi.fraunhofer.de/HM-doc/

[7]. D. T. Dang-Nguyen, C. Pasquini, V. Conotter, and G. Boato, "RAISE - A raw images dataset for digital image forensics," *AMC Multimedia Systems, Portland, Oregon*, March 18-20, 2015.

[8]. G. Bjontegaard, "Calculation of average PSNR differences between RD-curves," in *Proc. 13th VCEG Meeting*, pp. 1–5, Austin, TX, USA, Jan. 2001.

Chou-Chen Wang, et. al. "A Deep Learning Approach for HEVC Intra Coding." *The International Journal of Engineering and Science (IJES)*, 10(10), (2021): pp. 01-07.