Journal of Multimedia Information System VOL. 11, NO. 3, September 2024 (pp. 175-184): ISSN 2383-7632 (Online)
https://doi.org/10.33851/JMIS.2024.11.3.175

Deep Inter Prediction for Versatile Video Coding (VVC)

Qurat Ul Ain Aisha[1†], Young-Ju Choi[1†], Jongho Kim[2],
Sung-Chang Lim[2], Jin Soo Choi[2], Byung-Gyu Kim[1*]

Abstract

A sophisticated video surveillance system involves the problem of limited video storage to record video data for long time. Video compression technology is an effective solution to address this problem. Inspired by the success of neural network-based approaches in computer vision, research on neural network-based video coding has emerged. With the aim of achieving improved compression efficiency, an investigation on inter prediction plays a crucial role in neural network-based video coding. In this paper, we propose a convolutional neural network (CNN)-based generation and enhancement method for inter prediction (GEIP) in the Versatile Video Coding (VVC) standard. By leveraging fused features and self-attended features based on attention mechanism, the proposed method maximizes inter prediction performance. When compared with VTM-11.0 NNVC-1.0 anchor, it is verified that the BDrate reduction of the proposed method can be achieved up to 7.06% on Y component under random access (RA) configuration.

Key Words: Deep Learning, Inter Prediction, Frame Generation, Versatile Video Coding (VVC).

I. INTRODUCTION

A video surveillance system [1-13] is composed of video acquisition devices, data storage, and the system of data processing. Earlier video surveillance systems consisted of simple video acquisition. Recently, a video surveillance system has become a sophisticated automated system based on intelligent video analysis algorithms [13]. This system can be applied to various applications such as traffic control, accident prediction, crime prevention, motion detection, and homeland security [1]. However, a video surveillance system requires a large amount of storage. And it is inefficient to increase memory capacity continuously and take a large bandwidth when it is transmitted. Therefore, video compression technology is an effective solution to resolve the problem of restricted storage [14].

Following the successful standardization of advanced video coding (H.264/AVC) [15] and high-efficiency video coding (H.265/HEVC) [12], the joint video exports team (JVET) of ITU-T Video Coding Experts Group (VCEG) and ISO/IEC Moving Picture Experts Group (MPEG) has been finalized the Versatile Video Coding (VVC) [16] standard in July 2020. Designed with the intention of achieving a substantial reduction in bit rate compared to HEVC while maintaining the same level of visual quality, the VVC aims to provide improved compression efficiency. This standard is expected to be utilized in a variety of applications, including 8 K and higher resolution videos, game videos, screen content videos, 360-degree videos, high dynamic range (HDR) content, and adaptive resolution videos.

However, achieving such compression improvements requires high complexity, and for future standards beyond VVC, the development of new approaches and technologies will be necessary. Inspired by the success of the neural network-based approaches in the field of computer vision [17-18], neural network-based research has also started to emerge in video coding. Recently, various neural network-based coding techniques have been explored and validated through the exploration experiment (EE) of the JVET neural network-based video coding (NNVC) [19].

Research on in-loop filtering and post-filtering has been conducted actively and extensively. On the other hand, a few techniques related to prediction process have been proposed. The essence of video coding lies in eliminating the redundancy of signals. Intra and inter prediction processes

Manuscript received September 06, 2024; Revised September 18, 2024; Accepted September 20, 2024. (ID No. JMIS-24M-09-027)
Corresponding Author (*): Byung-Gyu Kim, +82-2-2077-7293, bg.kim@sookmyung.ac.kr
[†]Authors are equally contributed.
[1]Department of IT Engineering, Sookmyung Women's University, Seoul, Korea, E-mail: aisha.q@ivpl.sm.ac.kr, yj.choi@ivpl.sm.ac.kr, bg.kim@sookmyung.ac.kr
[2]ETRI, Daejeon, Korea, E-mail: pooney@etri.re.kr, sclim@etri.re.kr, jschoi@etri.re.kr

generate prediction block utilizing spatial signals around the current coding unit (CU) within a frame and temporal signals in previous and/or future neighboring frames, respectively. Therefore, the prediction process, which makes the most significant contribution to redundancy removal, has a substantial impact on the overall compression efficiency.

Especially, when it comes to video data, utilizing the similarity between frames is crucial. Therefore, improving the performance of the inter prediction process can greatly enhance the overall compression efficiency. The existing deep learning-based methods can be divided into two categories: enhancement of prediction block [2,11,20] and generation of bi-prediction block [10,21]. To enhance the uni-prediction module, the prediction block can be improved using neural network.

For the bi-prediction module, the neural network-based fusion of two blocks can be employed to generate a bi-prediction block, or the existing bi-prediction block can be enhanced applying a similar approach as uni-prediction. However, despite the importance of both generating and improving prediction blocks, existing research have enhanced the inter prediction process by focusing on one of these aspects. In this paper, we propose the convolutional neural network (CNN)-based generation and enhancement method for inter prediction (GEIP) in VVC.

The proposed method utilizes an attention mechanism to create the fused feature of two prediction signals in generation model and the selfattended feature of a prediction signal in enhance model. Different from the existing deep learning-based studies, the proposed framework contains both of the enhancement of uni-prediction block and the generation of bi-prediction block, allowing for maximizing the performance of inter prediction.

In the proposed method, the enhancement network is not applied to bi-prediction block where the generation network is employed to avoid the over-filtering problem. The rest of this paper is organized as follows. In Section 2, we describe the related works. In Section 3, we present the details of the proposed method. The experimental results are shown in Section 4. Finally, Section 5 makes the concluding remarks for this paper.

II. RELATED WORKS

2.1. Traditional Inter Prediction method in VVC

In VVC, several novel techniques have been proposed to enhance the inter-picture prediction features beyond HEVC [25]. In addition to traditional motion vector prediction (MVP) methods in HEVC, two novel types of MVP were introduced in VVC: history-based MVP (HMVP) and pairwise average MVP (PAMVP).

Furthermore, the VVC has adopted three additional merge modes: merge mode with motion vector differences (MMVD), geometric partitioning mode (GPM), and combined inter and intra prediction (CIIP). The affine motion compensation (AMC) was newly applied beyond the translation motion model-based motion compensation. The weighted prediction (WP) is used to compensate for the inter prediction signal and improve coding efficiency. The WP in VVC signals the weight and offset for each reference picture and compensates for each block. In the VVC standard, two new techniques were introduced for weighted bi-prediction at the coding unit (CU) level. The first method, called bi-prediction with CU-level weight (BCW), allows for weighted prediction at the CU level. The second method, known as refinement with bi-directional optical flow (BDOF), utilizes bi-directional optical flow for further enhancement.

2.2. Deep Learning Based Inter Prediction Method in Video Coding

With the successful application of deep learning in various computer vision tasks, numerous research leveraging deep learning techniques have emerged in the field of video coding. To generate the bi-prediction block in HEVC, Zhao et al. [21] introduced a CNN-based approach that employed a patch-to-patch inference strategy. Similarly, Mao and Yu [10] proposed a CNN-based bi-prediction method called STCNN, which leveraged spatial neighboring regions and temporal display orders as additional inputs to improve the accuracy of prediction.

These methods aim to substitute the averaging bi-prediction approach in HEVC standard.

For enhancement of the uni-prediction block in HEVC, Huo et al. [2] and Wang et al. [20] proposed the CNN based motion compensation refinement (CNNMCR) and the neural network-based inter prediction (NNIP), respectively. The objective of these methods is to enhance the prediction blocks generated from the conventional standard prediction approaches.

Recently, Merkle et al. [11] proposed a CNN-based method to enhance the motion compensated prediction signal of the inter prediction block by integrating spatial and temporal reference samples. Despite these attempts to improve the inter prediction, a framework that combines both the neural network-based generation and enhancement of prediction block has not been introduced. Both aspects are crucial in inter prediction, and the proposed framework maximizes the improvement of inter prediction performance. Bao et al. presented the joint reference frame synthesis (RFS) and post-processing filter enhancement (PFE) for Versatile Video Coding (VVC), aiming to explore the combination of different neural network-based video coding (NNVC) tools to better utilize the hierarchical bi-directional coding structure of VVC [23]. Both RFS and PFE

utilized the Space-Time Enhancement Network (STENet), which received two input frames with artifacts and produced two enhanced frames with suppressed artifacts, along with an intermediate synthesized frame. Also, Merkle et al. have suggested a scheme for improving the prediction signal of inter blocks with a residual CNN that incorporates spatial and temporal reference samples [24]. They introduced an additional signal plane with constrained spatial reference samples which enabled decoupling the network from the intra decoding loop.

Unlike the aforementioned methods, the proposed approach implements deep learning technique to calculate an optimal reference frame through generation or synthesis for bidirectional inter-frame prediction. Additionally, it utilizes an attention-based architecture to extract optimal features.

III. METHODOLOGY

In this section, the proposed CNN-based generation and enhancement method for inter prediction (GEIP) is presented. It begins by describing the architecture of the proposed networks. Then, it provides an overview of how the proposed networks are integrated into the VVC framework.

3.1. Architecture

The structure of the generation and enhancement networks in the proposed CNN-based inter prediction method is illustrated in Fig. 1. The proposed networks are designed utilizing the attention-based bi-prediction network (ABPN) [25] as a foundation. The proposed generation and enhancement networks share the same architecture as shown in Fig. 1. For the generation network, given two input reference prediction blocks I_{P0} and I_{P1}, the goal of the network is to generate the bi-prediction block. In the uni-prediction, we apply the enhancement network to the traditional uni-prediction block.

Since uni-prediction block is obtained by using only one temporal reference data, I_{P0} and I_{P1} are identical. In addition, the normalized slice quantization parameter (QP) map I_{QP} is utilized as an input to train and apply the same model for all QP values. Each of the two concatenated prediction inputs with QP map are fed into three ConvBlock to increase the feature dimensionality. In the proposed networks, the ConvBlock consists of 2D CNN layer and the leaky rectifier linear unit (LeakyReLU). The attention map between two input features is computed using the dot product and sigmoid activation function. Let F^i_{P0} and F^i_{P1} denote the output feature maps of the i-th ConvBlock for I_{P0} and I_{P1}, respectively. The attended features are computed as follows:

$$AF_{P0} = F^1_{P0} \otimes \sigma(F^3_{P0} \odot F^3_{P1}), \quad (1)$$

where $\odot$, σ, $\otimes$, AF_{P0}, and AF_{P1} denote the dot product,

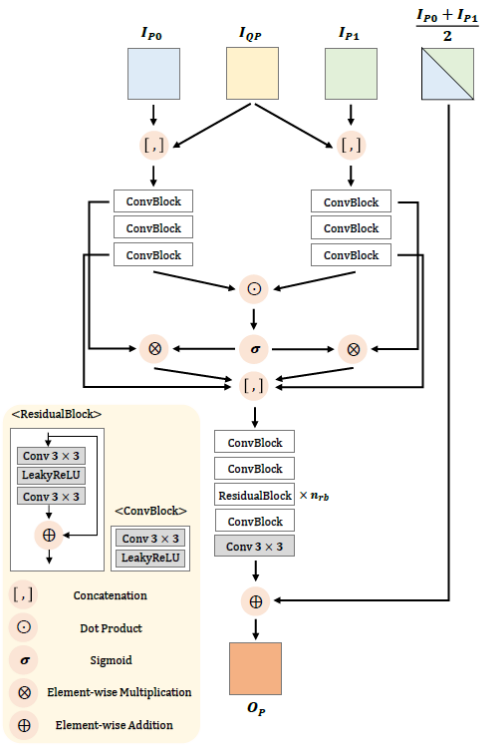


Fig. 1. The illustration of the proposed generation and enhancement networks for inter prediction.

$$AF_{P1} = F^1_{P1} \otimes \sigma(F^3_{P0} \odot F^3_{P1}), \quad (2)$$

the sigmoid activation function, the element-wise multiplication, the attended features for I_{P0} and I_{P1}, respectively.

Therefore, the attended features represent the fused feature between two prediction blocks and the self-attended feature of a prediction block in the generation and enhancement networks, respectively. In other words, same structure of attention module can be applied to different purpose in the proposed method.

3.2. Integration to VVC Reference Software

The proposed method is integrated into the motion compensation in VVC. Based on the experimental results demonstrated in ABPN [25], the proposed networks are applied only to CUs with sizes 128×128, 64×64, and 32×32.

Fig. 2 presents the flow-chart of the CNN-based generation and enhancement method for inter prediction (GEIP). First, uni-prediction process is performed for each reference picture.

In Fig. 2, idx_{ref}, $n_{refList}$, and $size_{cu}$ denote the current index of reference picture, the number of reference picture list, and the size of CU, respectively. Therefore, the 'Uni-prediction block' in Fig. 2 indicates the optimal uni-prediction block of traditional video coding at the current index of reference picture. The enhancement network is applied to the prediction block obtained from the traditional uni-pre-

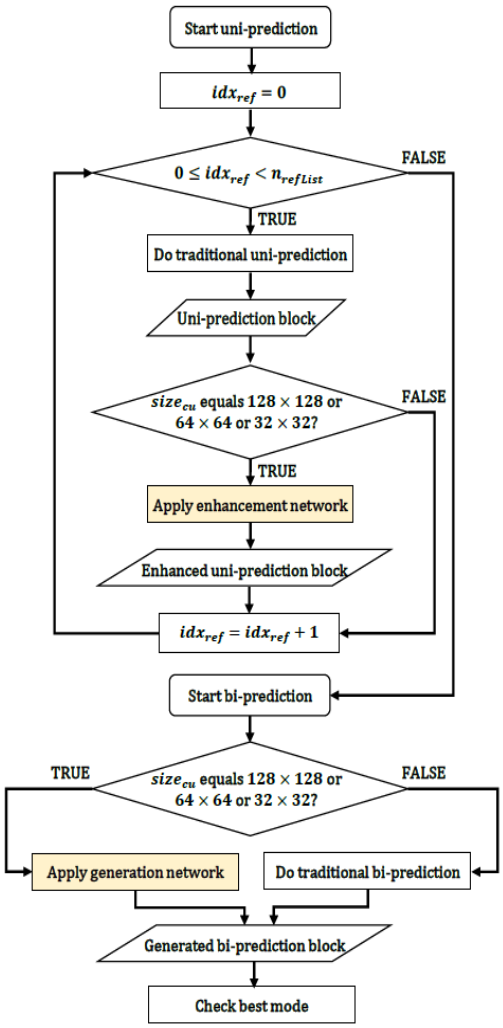


Fig. 2. The flowchart of the proposed CNN-based generation and enhancement method for inter prediction (GEIP) in the motion compensation.

diction process. In other words, we can change the existing uni-prediction block into the 'Enhanced uni-prediction block' by the proposed method.

Second, the proposed generation network replaces the traditional bi-prediction modules for three types of block. Therefore, the traditional bi-prediction methods in VTM such as the averaging mode, BCW, BDOF, WP are not performed if the proposed method is applied to the current CU. At this time, the decoder-side motion vector refinement (DMVR) can be adopted together with the proposed method.

Different from existing studies, the proposed framework encompasses both the enhancement of uni-prediction blocks and the generation of bi-prediction blocks, maximizing inter prediction performance. Furthermore, the proposed scheme has an advantage that it does not require additional flag bits. To avoid over-filtering issues, the enhancement network is not applied to the bi-prediction block where the generation network is employed.

IV. EXPERIMENT RESULTS

4.1. Experimental Setting

4.1.1. Network Training

We utilize the BVI-DVC sequence dataset [9] to generate the training dataset. The BVI-DVC dataset is the JVET common test conditions (CTC) training data for NNVC [4]. It contains 200 sequences, which are structured into four different resolutions. The VTM-11.0 NNVC-1.0 reference software is employed to compress the sequences with random access (RA) configuration using five QPs={22, 27, 32, 37, 42}. In decoding phase, the prediction blocks for three CU sizes 128×128, 64×64, and 32×32 are extracted as the inputs of the proposed networks. The ground-truth (GT) blocks are cropped from raw video frames. In order to train and apply a single model regardless of the QP values or CU sizes for each generation and enhancement models, we utilize the normalized slice QP map as an input and the copy and paste up-sampling technique. This approach involves copying and pasting the existing input block within a 128×128 patch to utilize all three block types together for training. The number of training data for generation network is 3,828,550, and for enhancement network, it is 3,585,057. In addition, we utilize the random horizontal flip and 90° rotation augmentation methods. We used the Adam optimizer [3] and the cosine annealing scheme [6]. The learning rate was set to 4×10^{-4}. And the size of mini-batch and the number of iterations were set to 128 and 600 K. The proposed ABPN was implemented in PyTorch on a PC with an Intel(R) Xeon(R) Gold 6256 CPU@3.60 GHz and a NVIDIA Quadro RTX 8000-48 GB GPU. The proposed network consists of 64 features for each CNN layer except the last layer. And the number of residual block n_{rb} was set to 10. The proposed network has 1,110,657 parameters.

4.1.2. Encoding Configration

The proposed method has been integrated into VTM-11.0 NNVC-1.0 reference software. The PyTorch library 1.7.1 is adopted to implement the proposed method. We follow the JVET CTC for neural network-based video coding technology [4] and utilize the random access (RA) configuration under five QPs={22, 27, 32, 37, 42}. In the experiments, we conducted 1 group of picture (GOP) test for class B and 1 second test for classes C and D on VVC CTC sequences due to the lack of time for experiments.

4.2. Comparison with VVC

The results of the BD-rate reduction and encoding/ decoding computational complexity compared to VTM-11.0 NNVC-1.0 anchor on RA for the Y component are reported in Table 1. It is observed that 1.12% BD-rate saving can be achieved on average.

Table 1. BD-rate reduction and encoding/decoding computational complexity compared to VTM-11.0 NNVC-1.0 anchor under RA configuration.

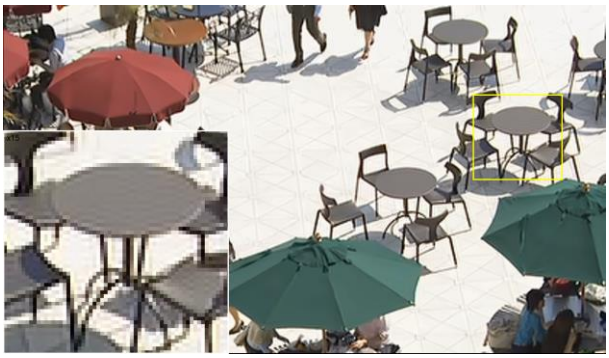
Class	Sequence	BD-rate (%)
Class B (1,920×1,080)	MarketPlace	−0.56
	RitualDance	0.36
	Cactus	−1.32
	BasketballDrive	−0.48
	BQTerrace	−1.20
Class C (832×480)	BasketballDrill	−0.77
	BQMall	−0.99
	PartyScene	−1.29
	RaceHorses	0.08
Class D (416×240)	BasketballPass	−0.61
	BQSquare	−7.06
	BlowingBubbles	−0.65
	RaceHorses	−0.09
Overall	−1.12	
Encoding running time	6,248	
Decoding running time	1,467	

In particular, the proposed method achieves up to 7.06% BD-rate reduction for *BQSquare*. As a result, the proposed method offers the potential to significantly improve inter prediction performance in video coding. Generally, the VVC standard considers only the format of the coded bitstream, syntax, and the operation of the decoder.

Therefore, the crucial measure for the computational complexity is the decoding running time. The proposed method showed 1,467% of decoding time on average under RA. Several recent JVET standard input documents for CNN-based inter prediction technology, JVETY0090 [7], JVET-Z0074 [8], and JVET-AA0082 [5], show 23,785%, 28,226%, and 175,647% of decoding time on average of classes B, C, and D under RA compared to VTM-11.0 NNVC-1.0 anchor. It can be seen that the proposed method is superior to the recent CNN-based inter prediction methods in terms of the computational complexity. Fig. 3 presents the qualitative results on *BQSquare* and *PartyScene*. It can be observed that the proposed method can remove the noise on the table by comparing Fig. 3(a) and Fig. 3(b). Therefore, we can verify that the proposed method can reduce the compression artifacts.

As shown in Figs 3(c) and (d) the proposed GEIP more accurately restores the texture of clothes of moving person. It means that the proposed method recovers more clear edge details.

Fig. 4 shows the visualization results of the *InterDir* flag for CUs in *BQSquare* sequence. For this figure, we used two frames (POC 38 and 48) under QP 27 and RA configuration. The blue, green, and red blocks indicate L0 unidi-



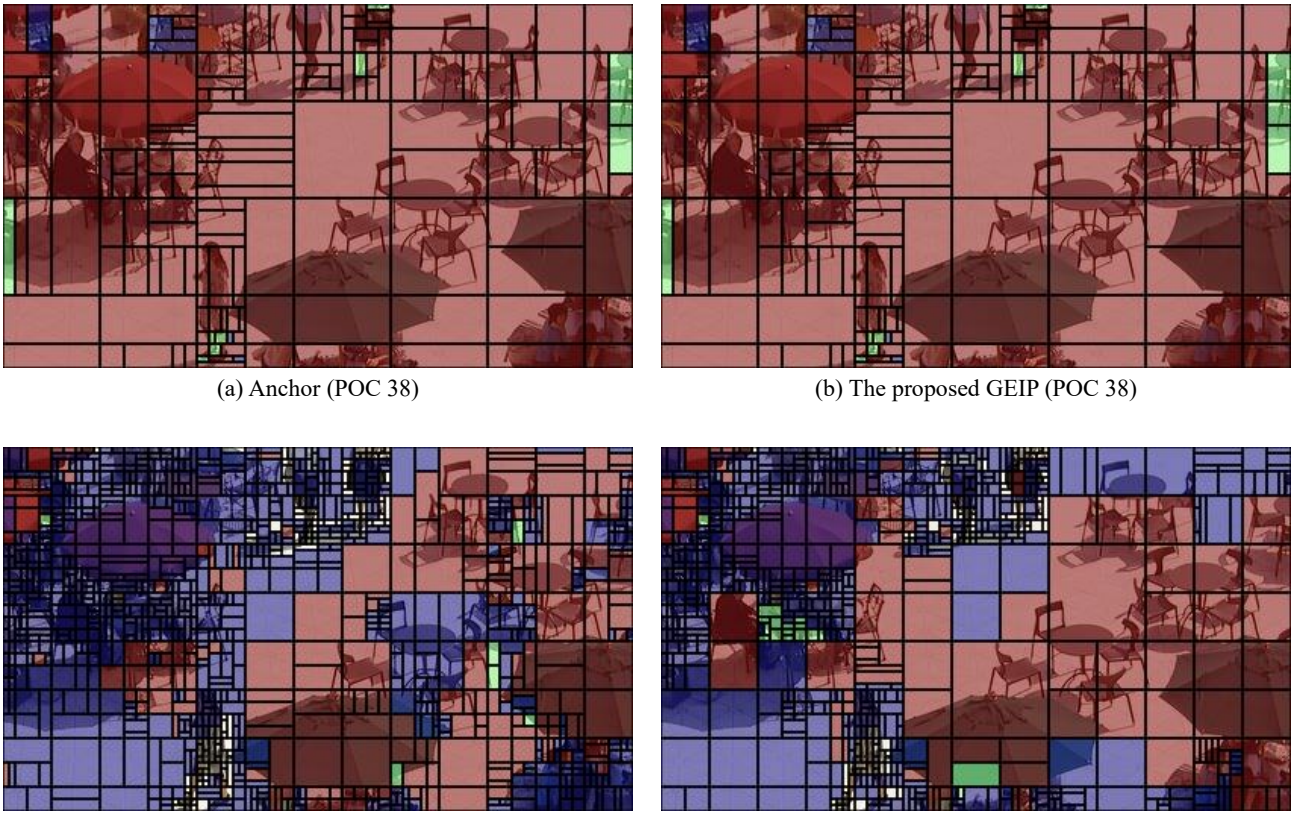
(a) Anchor (POC 12 of BQSquare under QP=27)

(b) The proposed GEIP (POC 12 of BQSquare under QP=27)

(c) Anchor (POC 7 of *PartyScene* under QP=37)

(d) The proposed GEIP (POC 7 of *PartyScene* under QP=37)

Fig. 3. Visual quality improvement of the proposed GEIP compared to VTM-11.0 NNVC-1.0 anchor under RA configuration.



(a) Anchor (POC 38) (b) The proposed GEIP (POC 38)

(c) Anchor (POC 48) (d) The proposed GEIP (POC 48)

Fig. 4. Illustration of the *InterDir* flag for CUs in *BQSquare* compared to VTM-11.0 NNVC-1.0 anchor under QP=27 and RA configuration.

rectional, L1 uni-directional, and bi-directional *InterDir* flags, respectively. On the whole, the portion of three CU sizes 128×128, 64×64, and 32×32 is increased compared to VTM anchor. Since the small size of blocks can be merged into the large size of blocks, coding bits for a CU can be reduced.

V. CONCLUSION

In this paper, we have proposed a convolutional neural network (CNN)-based generation and enhancement method for inter prediction in the Versatile Video Coding (VVC) standard. By utilizing an attention mechanism, the proposed method achieves both the enhancement of uni-prediction blocks and the generation of bi-prediction blocks, maximizing the performance of inter prediction. Different from existing research, the proposed framework addressed both aspects of generating and improving prediction blocks, offering a comprehensive approach for inter prediction. The experimental results demonstrated the effectiveness of the proposed method in improving compression efficiency. The proposed approach achieved up to 7.06% BD-rate saving for the Y component under RA compared with the VTM-11.0 NNVC-1.0 anchor. As a result, the proposed method can contribute to the advancement of neural network-based video coding techniques. In future work, the computational complexity of the proposed inter prediction scheme can be reduced by redesigning the algorithm of integration.

ACKNOWLEDGMENT

This work was supported by Institute for Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2017-0-00072, Development of Audio/Video Coding and Light Field Media Fundamental Technologies for Ultra Realistic Tera-media).

REFERENCES

[1] O. Elharrouss, N. Almaadeed, and S. Al-Maadeed, "A review of video surveillance systems," *Journal of Visual Communication and Image Representation*, vol. 77, p. 103116, Jan. 2021.

[2] S. Huo, D. Liu, F. Wu, and H. Li, "Convolutional neural network-based motion compensation refinement for video coding," in *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*, IEEE, Jan. 2018, pp. 1-4.

[3] D. Kingma and J. B. Adam, "A method for stochastic

optimization," in *International Conference on Learning Representations (ICLR)*, Apr. 2015, pp. 1-15.

[4] S. Liu, A. Segall, E. Alshina, and R. L. Liao. "JVET common test conditions and evaluation procedures for neural network based video coding technology," in *Document JVET-X2016, 24th JVET Meeting*, Apr. 2021, pp. 1-10.

[5] Z. Liu, X. Xu, S. Liu, J. Jia, and Z. Chen, "AHG11: Deep reference frame generation for VVC inter prediction enhancement," in *Document JVET-AA0082, 27th JVET Meeting*, May 2022, pp. 1-7.

[6] I. Loshchilov and F. Hutter, "Sgdr: Stochastic gradient descent with warm restarts," *arXiv Prep. arXiv: 1608. 03983*, pp. 1-16, Apr. 2016.

[7] C. Ma, R. L. Liao, and Y. Ye., "AHG11: Neural network-based motion compensation enhancement for video coding," in *Document JVET-Y0090, 25th JVET Meeting*, May 2022, pp. 1-4.

[8] C. Ma, R. L. Liao, and Y. Ye, "AHG11: Neural network based motion compensation enhancement for video coding," in *Document JVET-Z0074, 26th JVET Meeting*, May 2022, pp. 1-5.

[9] D. Ma, F. Zhang, and D. R. Bull, "BVI-DVC: A training database for deep video compression," *IEEE Tran-sactions on Multimedia*, vol. 24, pp. 3847-3858, Apr. 2021.

[10] J. Mao and L. Yu, "Convolutional neural network based biprediction utilizing spatial and temporal information in video coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 7, pp. 1856-1870, Jan. 2019

[11] P. Merkle, M. Winken, J. Pfaff, H. Schwarz, D. Marpe, and T. Wiegand, "Intra-inter prediction for versatile video coding using a residual convolutional neural network," in *2022 IEEE International Conference on Image Processing (ICIP)*, IEEE, Jan. 2022, pp. 1711-1715.

[12] G. J. Sullivan, J. R. Ohm, W. J. Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1649-1668, Jan. 2012.

[13] V. Tsakanikas and T. Dagiuklas, "Video surveillance systems current status and future trends," *Computers & Electrical Engineering*, vol. 70, pp. 736-753, Jan. 2018.

[14] L. Zonglei and X. Xianhong, "Deep compression: A compression technology for apron surveillance video," *IEEE Access*, vol. 7, pp. 129966-129974, Jan 2019.

[15] T. Wiegand, G. J. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the H. 264/avc video coding standard," IEEE Transactions on Circuits and Systems for Video Technology, vol. 13, no. 7, pp. 560-576, Jan. 2003

[16] B. Bross, Y. K. Wang, Y. Ye, S. Liu, J. Chen, and G. J. Sullivan, et al., "Overview of the versatile video coding (VVC) standard and its applications," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 10, pp. 3736-3764, Jan. 2021.

[17] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-re-solution," in *Computer Vision–ECCV 2014: 13th European Conference*, Zurich, Switzerland, Jan. 2014, pp. 184-199.

[18] Y. S. Chen, Y. C. Wang, M. H. Kao, and Y. Y. Chuang, "Deep photo enhancer: Unpaired learning for image enhancement from photographs with gans," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Jan. 2018, pp. 6306-6314.

[19] E. Alshina, F. Galpin, Y. Li, D. Rusanovskyy, M. Santamaria, and J. Ström, L. et al., "Explo-ration experiments on neural network-based video coding (EE1)," in *document JVET-AD2023, 30th JVET Mee-ting*, Jan. 2023. pp. 1-12.

[20] 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, Jan. 2018, pp. 1-6.

[21] Z. Zhao, S. Wang, S. Wang, X. Zhang, S. Ma, and J. Yang, "Enhanced bi-prediction with convolutional neural network for high-efficiency video coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 11, pp. 3291-3301, Jan. 2018.

[22] J. Chen, Y. Ye, and S. H. Kim, "Algorithm description for versatile video coding and test model 11 (VTM 11)," in *Document JVET-T2002, 20th JVET Meeting*, Feb. 2020, pp. 1-101.

[23] W. Bao, Y. Zhang, J. Jia, Z. Chen, and S. Liu, "Joint reference frame synthesis and post filter enhancement for versatile video coding," *https://arxiv.org/pdf/2404.18058*, 2024.

[24] P. Merkle, M. Winken, J. Pfaff, H. Schwarz, D. Marpe and T. Wiegand, "Spatio-temporal convolutional neural network for enhanced inter prediction in video coding," in *IEEE Transactions on Image Processing*, 2024, vol. 33, pp. 4738-4752.

[25] Y. J. Choi, Y. W. Lee, J. Kim, S. Y. Jeong, J. S. Choi, and B. G. Kim, "Attention-based bi-prediction network for versatile video coding (VVC) over 5g network," *Sensors*, vol. 23, no. 5, p. 2631, Feb. 2023.

AUTHORS

Qurat Ul Ain Aisha completed her Bachelor's degree in Software Engineering in 2022 from Bahria University, Islamabad, Pakistan. Currently, she is pursuing an integrated degree at Sookmyung Women's University, Korea. Her research interests lie in the fields of video coding standards and deep learning techniques, where her aim is to explore innovative solutions for enhancing video compression and quality.
With a strong foundation in software engineering and a growing expertise in advanced machine learning methods, she is passionate about applying advance technologies to real-world multimedia challenges.

Young-Ju Choi received her B.S. degree in the Department of Information Technology (IT) Engineering from Sookmyung Women's University, Korea, in 2017, and the Ph.D. degree in the Department of Information Technology (IT) Engineering from Sookmyung Women's University, South Korea, in 2024. Her research interests include video super-resolution (VSR), video coding standard, with the focus on deep-learning-based methods.

Jongho Kim received B.S. degree from Control and Computer Engineering Department, Korea Maritime University in 2005 and he also received M.S. degree from University of Science and Technology (UST) in 2007. He joined Electronics and Telecommunications Research Institute (ETRI) in Daejeon, Korea in 2008 and has been a senior researcher since 2016. He is currently pursuing his Ph.D. degree at Korea Advanced Institute of Science and Technology (KAIST). His research interests include video standardization activities of ITU-T VCEG and ISO/IEC MPEG, perceptual video coding, and machine learning based compression.

Sung-Chang Lim received his B.S. (with highest honors), M.S., and Ph.D. degrees in Computer Engineering from Sejong University, Seoul, Korea, in 2006, 2008, and 2022, respectively. In 2008, he joined Electronics and Telecommunications Research Institute (ETRI), Daejeon, Korea, where he is currently a Principal Researcher at Telecommunication & Media Research Laboratory. Since 2005, he has been involved in the development of several international video coding standards, including AVC FRExt, MVC, HEVC, and VVC, with more than 80 technical contributions. He has authored and co-authored over 50 academic publications and holds numerous internationally issued patents and patent applications in the field of video coding. His research interests include video coding and image processing.

Jin Soo Choi received the B.E., M.E., and Ph.D. dgrees in electronic engineering from Kyungpook National University, Korea, in 1990, 1992, and 1996, respectively. Since 1996, he has been a principal member of engineering staff in ETRI, Korea. He has been involved in developing the MPEG-4 codec system, data broadcasting system, and UDTV. His research interests include visual signal processing and interactive services in the field of the digital broadcasting technology.

Byung-Gyu Kim has received his B.S. degree from Pusan National University, Korea, in 1996 and an M.S. degree from Korea Advanced Institute of Science and Technology (KAIST) in 1998. In 2004, he received a Ph.D. degree in the Department of Electrical Engineering and Computer Science from Korea Advanced Institute of Science and Technology (KAIST). In March 2004, he joined in the real-time multimedia research team at the Electronics and Telecommunications Research Institute (ETRI), Korea where he was a senior researcher. In February 2009, he joined the Division of Computer Science and Engineering at SunMoon University, Korea where he is currently a professor. He has been serving as a Professional Reviewer in many academic journals, including IEEE, ACM, Elsevier, Springer, Oxford, SPIE, IET, MDPI, IT&T, and so on. He has been serving as an Associate Editor for *Circuits, Systems and Signal Processing* (Springer), *The Journal of Supercomputing* (Springer), *The Journal of Real-Time Image Processing* (Springer), *Journal of Imaging Science and Technology* (IS&T), *Heliyon-Computer Science* (Cell press), *CAAI Transactions on Intelligence Technology* (IET), and *Applied Sciences* (MDPI). Since March 2018, he has been serving as the Editor-in-Chief for *The Journal of Multimedia Information System* and an Associate Editor for *IEEE ACCESS Journal.* He has received the Special Merit Award for Outstanding Paper from the IEEE Consumer Electronics Society, at IEEE ICCE 2012, the Certification Appreciation Award from the SPIE Optical Engineering, in 2013. He has been honored as an IEEE Senior Member, in 2015. He has also received the Excellent Paper Award from the IEEE Consumer Electronics Society, at IEEE ICCE 2021.

He has published over 270 international journal and conference papers, patents in his field. His research interests include image and video object segmentation for the content-based image coding, video coding techniques, 3D video signal processing, and intelligent information system for image signal processing. He is a senior member of IEEE and a professional member of ACM, and IEICE.