Attention Based Deep Reference Frame Enhancement for VVC Inter Prediction

Linchen Xu*, Zhikai Liu* and Fan Liang*§
*Sun Yat-sen University, China
E-mail:{xulch3,liuzhk6}@mail2.sysu.edu.cn,
§Corresponding author: isslf@mail.sysu.edu.cn

***Abstract*—This paper proposes a deep-learning based inter-frame coding enhancement method that significantly improves bidirectional prediction performance under the H.266/VVC standard through the integration of diverse attention mechanisms. Our method follows the deep reference frame (DRF) generation approach, while addressing the limitations of existing methods. To improve quality of the generated DRF, we introduce two novel modules: an inter-frame attention module which handles the insufficient utilization of high-level features in optical flow estimation, and a spatial-frequency attention based module that compensates the lack of frequency-domain modeling in frame synthesis. Comprehensive validation on the VTM-15.0 platform demonstrates that our framework achieves average BD-rate savings of 4.20%/8.53%/9.34% for Y/U/V components respectively under RA configuration. Ablation studies further confirm the efficacy of each proposed module. Compared with existing DRF coding schemes, our method maintains comparable performance while reducing computational complexity, thereby providing a practical solution for neural network-based video coding applications.**

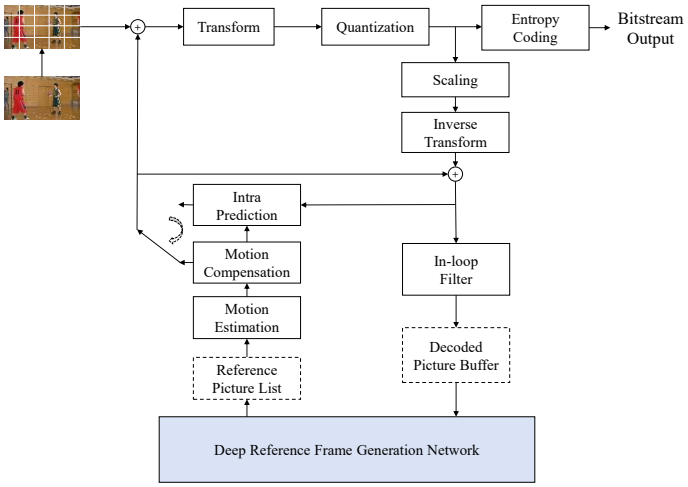


Fig. 1. Framework of Deep Reference Frame generation method inside traditional encoder.

I. Introduction

Video coding standards in the past decades have consequently evolved to address bandwidth constraints introduced by data storage and transmission, progressing from H.264/AVC [1] to the more efficient H.265/HEVC [2]. The latest H.266/VVC [3] standard represents a significant breakthrough, achieving approximately 50% bitrate reduction compared to its predecessor while preserving equivalent visual quality.

In recent years, deep learning-based video coding techniques, with their potential in feature extraction and nonlinear modeling, have become as a crucial research area. One way of utilize neural networks is to build end-to-end video coding networks. Examples include the pioneer deep video compression framework (DVC) [4], or subsequent deep contextual video compression (DCVC) [5]. Nevertheless, such neural video compression based approach usually cause complexity and is not hardware friendly. Another way is to incorporate neural networks into traditional coding frameworks, and it exploits advantages of traditional block-based coding methods. The Neural Network-based Video Coding (NNVC) group, established by the Joint Video Experts Team (JVET) in 2020, has been actively exploring such approaches. The proposals discussed include intra prediction, inter-prediction, in-loop filtering, post filtering and super-resolution. These emerging methodologies have demonstrated significant potential in improving coding efficiency.

To utilize deep learning to enhance the performance of inter-frame prediction coding, there are mainly two approaches in general, i.e., based on block and frame respectively. The first approach targets motion compensation. Studies [6] and [7] both use Convolution Neural Networks (CNN) to enhance prediction blocks. Wang et al. [8] used reconstructed information to predict the residual required for motion compensation, so as to reduce transmitted residual information. Merkle et al. [9] optimized the above methods in terms of coding implementation.

The frame-based approach uses deep neural networks to optimizes the quality of the reference frames. There are studies [10], [11] used deep learning to achieve high-precision sub-pixel interpolation, while recent studies trend to generate an virtual non-existence reference frame, also called the deep reference frame (DRF), that better correlates the current frame. Its structure is shown in Fig 1. Due to the similarity between DRF generation and video frame interpolation (VFI), the designs in latter study have significant value for reference. Kernel-based methods [12] in VFI utilize learned kernels to estimate motion. Inspired by such method, studies [13], [14], [15], [16] use kernel-based motion estimation as part of deep

979-8-3315-7206-8/25/$31.00

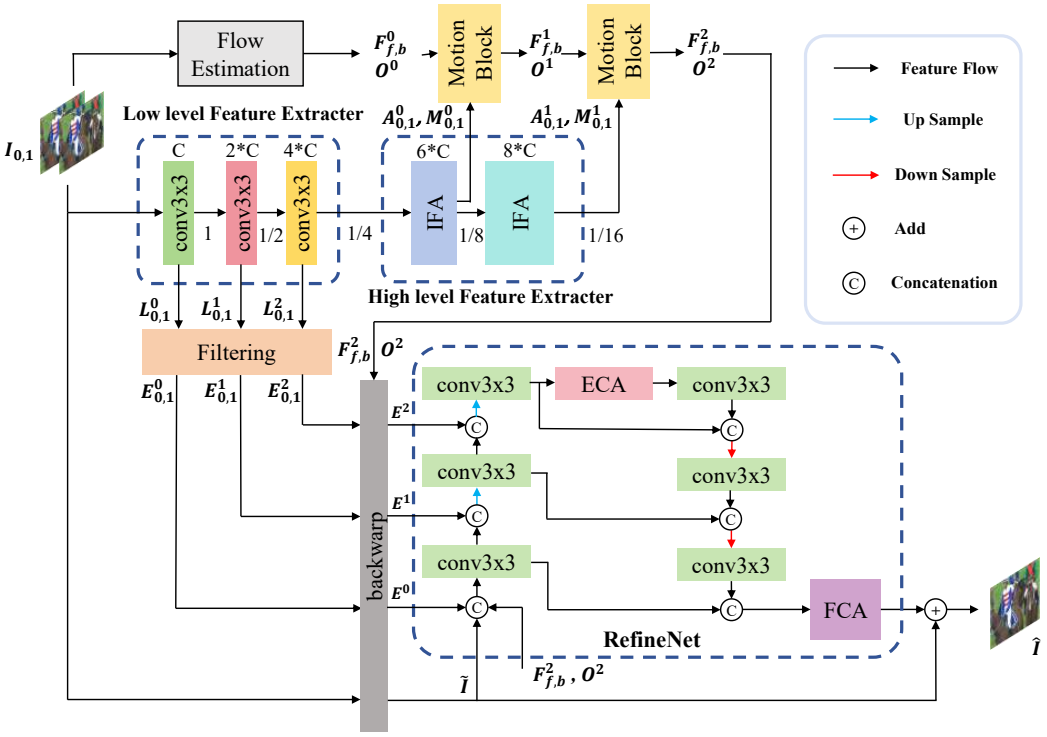


Fig. 2. Overall Structure of the proposed attention-based deep reference frame generation method.

neural networks to generate and enhance virtual reference frames in inter-prediction, although it brings limited improvements in coding efficiency compared to its parameterize and computational overhead. Another popular VFI method is based on optical flow estimation[17], [18]. In recent years, studies have utilized 3D convolution[19], multi-scale approach[20], knowledge distillation [21], correlation calculation[22], cross-attention[23] to form optical flow, and its accuracy attains remarkable results in recent years. And there are also flow and kernel combined methods[24], [25] but have similar problems when applied in DRF generation.

The strong temporal modeling of optical flow enhances intermediate frame synthesis accuracy, thus offers an effective design paradigm for DRF generation. Hu et al.[26] proposed an error-corrected auto-regression network to perform VFI while maintaining information progression. Jia at al. [27] utilize an model inspired by [12] aiming at inter-prediction in H.266/VVC, although later study [28] which use [20] as core motion estimation achieve better performance with lower complexity. Notably, the journal version [29] of study [28] proposed an prediction and enhancement DRF generation model, achieving state-of-the-art performance among existing approaches. Furthermore, for practical employments, studies [30], [31] explore lightweight DRF models, with corresponding degradation in coding efficiency compared to [28].

The application of neural network tools in inter-frame coding motivates the adoption of more powerful architectures. Since Transformers and cross-attention have proven effective at capturing spatio-temporal feature relationships in visual tasks, they are well-suited for enhancing deep reference frames by accurately modeling dependencies across reference frames. Therefore, this paper proposes a deep reference frame method specifically designed for bi-directional inter-frame predictive coding. Our approach employs inter-frame attention to high level feature, enhance both features and optical flow. Furthermore, spatial and frequency-domain channel attention are introduced in order to generate more refined predicted frame. The proposed method has been integrated into VTM-15.0, achieving substantial performance improvements and demonstrating the effectiveness of our attention based DRF enhancement strategy.

II. Proposed Method

The overall framework of our deep reference frame generation is shown in Fig 2. The encoder selects two specific reconstructed frames in the DPB (Decoded Picture Buffer) in the inter prediction process, following the selecting strategy in study [29]. The generated frame will be inserted into both Reference Picture Lists (RPL0 and RPL1) with identical POC of current frame and the RPL index of 1, then encoder treats the DRF as common reference frame to perform block-based Rate-distortion Optimization (RDO).

A. Inter-frame Attention Based Enhancement

The proposed flow enhancement module addresses the limitations of existing optical flow estimation in DRF generation by leveraging high-level semantic features through the inter-frame attention mechanism. As illustrated in Fig. 2, the pro-

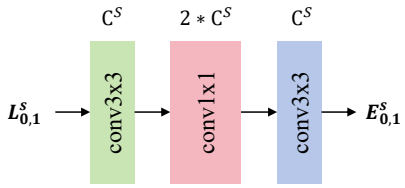


Fig. 3. Filtering Module in the proposed model.

cessing pipeline begins with a flow estimation network that generates initial forward and backward flow estimation $\boldsymbol{F}^0_{f,b}$ and corresponding occlusion mask $\boldsymbol{O}^0$ from input frames $\boldsymbol{I}_{0,1}$, as:

$$\boldsymbol{F}^0_{f,b}, \boldsymbol{O}^0 = \mathcal{F}(\boldsymbol{I}_{0,1}) \tag{1}$$

The input frames are also sent to a multi-scale feature extraction network to produce low-level spatial features $\boldsymbol{L}^s_{0,1}$ ($s = \{0, 1, 2\}$) at different resolutions. Scale factor compared to original inputs and channels at each stage are shown in Fig. 2, and basic channel is set to $\boldsymbol{C} = 16$. The core innovation lies in the inter-frame attention mechanism that computes cross-frame correlations. Following the design of [23], high-level features $\boldsymbol{H}^k_{0,1}$ ($k = \{0, 1\}$) high-level features are generated from $\boldsymbol{L}^s_{0,1}$ ($s = \{0, 1, 2\}$), and are windowed as $\Sigma\boldsymbol{B}^k_i$ and $\Sigma\boldsymbol{B}^k_j$. For each window ($i, j \in \{0, 1\}, i \neq j$), we have:

$$\boldsymbol{S} = \text{softmax}\left(\frac{\boldsymbol{Q}_i \boldsymbol{K}^T_j}{\sqrt{d}}\right)\boldsymbol{V}_j \tag{2}$$

where $\boldsymbol{Q} = \boldsymbol{B}_i\boldsymbol{W}_q$, $\boldsymbol{K} = \boldsymbol{B}_j\boldsymbol{W}_k$, and $\boldsymbol{V} = \boldsymbol{B}_j\boldsymbol{W}_v$ are learned projections of the high-level features. $\boldsymbol{S}$ will be used to derive appearance features $\boldsymbol{A}^k_{0,1}$ and motion features $\boldsymbol{M}^k_{0,1}$ ($k = \{0, 1\}$).

The flow refinement is achieved through two consecutive Motion Blocks $\mathcal{M}_0$ and $\mathcal{M}_1$ follow the structure in [23]. Each block processes the concatenation of input flow and attention features. The refinement process is formulated as:

$$\boldsymbol{F}^{n+1}_{f,b}, \boldsymbol{O}^{n+1} = \mathcal{M}_n(\boldsymbol{F}^n_{f,b}, \boldsymbol{O}^n, \boldsymbol{A}^n_{0,1}, \boldsymbol{M}^n_{0,1}), n \in \{0, 1\} \tag{3}$$

The final enhanced flow $\boldsymbol{F}^2_{f,b}$ and mask $\boldsymbol{O}^2$ demonstrates improved accuracy for both large displacements and fine details by effectively combining multi-scale attention contexts with the initial flow estimation. Then, an initial predicted frame can be obtained, as:

$$\tilde{\boldsymbol{I}} = \boldsymbol{O}^2 \odot \text{BW}(\boldsymbol{F}^2_f, \boldsymbol{I}_0) + (1 - \boldsymbol{O}^2) \odot \text{BW}(\boldsymbol{F}^2_b, \boldsymbol{I}_1) \tag{4}$$

where BW represents back-warp flow operation and $\odot$ denotes the Hadamard product.

B. Multi-scale Feature Filtering

In order to suppress artifacts caused by compression while preserving structural details in the input reference frame, we design a filtering module $\mathcal{G}$ for each scale. As shown in Fig. 3, it consists of three CNN layers, with two outer 3x3 convolution layers and 1x1 convolution layer as mapping layer [32] in the middle, which also has double amount of channels. The module processes the initial multi-scale low-level features $\boldsymbol{L}^s_{0,1}$ to obtain $\boldsymbol{E}^s_{0,1}$ as:

$$\boldsymbol{E}^s_{0,1} = \mathcal{G}^s(\boldsymbol{L}^s_{0,1}), s \in \{0, 1, 2\} \tag{5}$$

Then, the multi-scale filtered features of individual frames $\boldsymbol{E}^s_{0,1}$ are back-warped using the enhanced flow $\boldsymbol{F}^2_{f,b}$ to get synthesis feature $\boldsymbol{E}^s$, which will form the final predicted frame. The back-warping process here, denoted as BWS, use no occlusion mask (or equally, setting its value to 0.5) and performs necessary down-sample to $\boldsymbol{F}^2_{f,b}$ in order to match the scales, as:

$$\boldsymbol{E}^s = \text{BWS}(\boldsymbol{F}^2_{f,b}, \boldsymbol{E}^s_{0,1}), s \in \{0, 1, 2\} \tag{6}$$

C. Spatial and Frequency Attention Refinement

After obtaining multi-scale synthetic features, the final step involves performing multi-scale refinement to synthesize the output frame. The refinement network consists of a simple U-Net architecture [33] augmented with Spatial and Frequency Attention modules. To facilitate comprehensive interaction of multi-dimensional information in the spatial domain while avoiding dimensional reduction during this process, we introduce an Efficient Channel Attention (ECA) [34] module-based ResBlock at the lowest level of the U-Net. Following the U-Net processing, a Frequency Channel Attention (FCA) module [35] is further incorporated to enhance frequency-domain information, thereby strengthening the texture details of the synthesized frame. The complete workflow is illustrated as:

$$\hat{\boldsymbol{I}} = \tilde{\boldsymbol{I}} + \mathcal{R}(\boldsymbol{E}^s_{0,1}, \tilde{\boldsymbol{I}}, \boldsymbol{F}^2_{f,b}, \boldsymbol{O}^2), s \in \{0, 1, 2\} \tag{7}$$

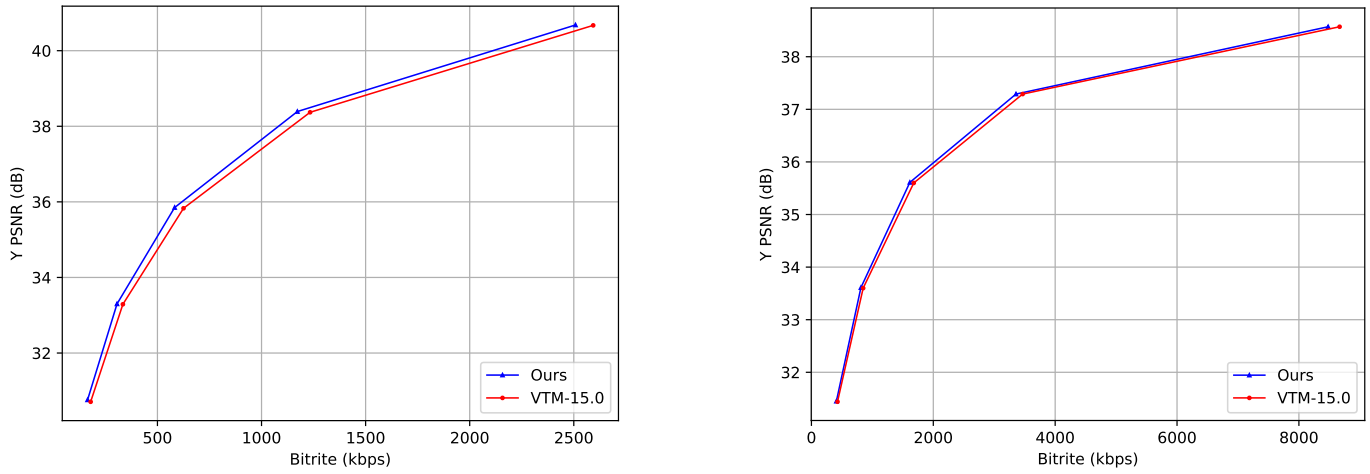
where the final generated frame $\hat{\boldsymbol{I}}$ is form by the initial predicted frame and a refined residual.

III. Experiment

A. Training Details

1) Training Data: The dataset employed in this study consists of compressed Vimeo-90K triplets. The original dataset comprises 91,701 triplets, with each triplet containing three consecutive images at a resolution of 448×256. To better simulate scenarios inside the encoder, we first encoded the images in the dataset using a VTM encoder to simulate reconstructed frames after compression. Each image within the triplets was compressed under the All-Intra (AI) configuration, with quantization parameters (QP) randomly selected from the set 22, 27, 32, 37, 42.

2) Training Strategy: The network was implemented using the PyTorch framework, with an L1 loss function and optimized using the Adam optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.99$). A pretrained small IFRNet [20] was utilized as the optical flow estimator, with its learning rate set to 10^{-5}, while the remaining parts of the network were trained with a learning rate of 10^{-4}. During training, for each triplet, the first and third images (compressed) served as network inputs, and the



(a) BQMall (b) Cactus

Fig. 4. RD-curve of two selected sequences (BQMall from Class C and Cactus from Class B). The red indicates results of VTM-15.0 anchor while the blue one represents the coding performance of our proposed method.

second image (uncompressed) was used as the ground truth. The entire network was trained on two TITAN RTX GPUs with a batch size of 24 for 200 epochs.

B. Experimental results

The proposed method was integrated into the VVC reference software VTM-15.0. Experiments were conducted following the JVET Common Test Conditions (CTC) for neural network-based video coding [36]. Table I presents the experimental results under the Random Access (RA) configuration. Coding performance improvements were evaluated using BD-rate (negative values indicate bitrate savings at the same PSNR), and the anchor is VTM-15.0. Each test sequence was encoded at five QP values (22, 27, 32, 37, 42), and the average BD-rate is reported. As shown in the table, the proposed method achieves BD-rate savings of -4.20%, -8.53%, -9.34% for Y, U and V component, respectively. The Rate-distortion curve on two selected sequence are also shown in Fig. 4, which intuitively demonstrate our model's superior performance over the anchor.

TABLE I
BD-RATE PERFORMANCE OF THE PROPOSED METHOD OVER VTM-15.0

Class	Sequence	BD-Rate		
		Y	U	V
ClassA1	Tango2	-4.05%	-6.11%	-7.95%
	FoodMarket4	-1.09%	-2.56%	-3.44%
	Campfire	-0.38%	-1.46%	-2.12%
	Average	-1.84%	-2.82%	-3.99%
ClassA2	CatRobot	-4.56%	-7.91%	-9.91%
	DaylightRoad2	-4.67%	-7.91%	-9.36%
	ParkRunning3	-1.37%	-3.44%	-4.88%
	Average	-3.53%	-7.07%	-7.09%
ClassB	MarketPlace	-2.43%	-7.33%	-6.52%
	RitualDance	-2.49%	-5.77%	-6.56%
	Cactus	-4.08%	-7.52%	-6.91%
	BasketballDrive	-3.33%	-9.06%	-9.36%
	BQTerrace	-2.53%	-8.01%	-6.88%
	Average	-2.97%	-7.54%	-7.25%
ClassC	RaceHorses	-4.57%	-10.19%	-10.41%
	BQMall	-7.17%	-14.17%	-14.47%
	PartyScene	-5.40%	-11.04%	-11.57%
	BasketballDrill	-4.88%	-11.52%	-12.74%
	Average	-5.51%	-11.73%	-12.30%
ClassD	RaceHorses	-6.58%	-15.32%	-16.35%
	BQSquare	-10.83%	-13.97%	-21.59%
	BlowingBubbles	-4.65%	-9.47%	-10.37%
	BasketballPass	-6.61%	-15.14%	-16.02%
	Average	-7.17%	-13.47%	-16.08%
Overall		**-4.20%**	**-8.53%**	**-9.34%**

C. Comparison with other works

We conducted comparative experiments with two existing interpolation-based methods under the RA configuration, both of which generate intermediate reference frames based on the VTM-15.0 framework. Among them, Hu et al. [26] proposed an error-correction autoregressive network, while Jia et al. [27] employed a 3D convolutional network for reference frame generation. Table II compares the BD-rate performance improvements of these methods. Since [26] was only tested at four QP values (22, 27, 32, 37), the same encoding settings were adopted for a fair comparison. The experimental data demonstrate that the proposed method achieves significant improvements across all test categories. On average, it delivers a 3.94% BD-rate reduction for the Y component, substantially outperforming [26] (1.54%) and [27] (1.70%). Moreover, the proposed method exhibits notable advantages in U/V components, particularly excelling in Class C and Class D sequences.

Table IV presents a comparative evaluation of the proposed method with two reference studies [27] and [29], focusing on their average BD-rate reduction and computational com-

TABLE II
BD-RATE PERFORMANCE COMPARED WITH OTHER WORKS

Class	BD-rate								
	Hu et al.[26]			Jia et al.[27]			Ours		
	Y	U	V	Y	U	V	Y	U	V
Class A1	-0.74%	-1.85%	-2.35%	-1.30%	-3.58%	-3.91%	-1.59%	-2.70%	-3.97%
Class A2	-0.70%	-3.23%	-2.52%	-1.78%	-6.26%	-4.54%	-3.16%	-6.84%	-7.02%
Class B	-0.81%	-2.88%	-3.33%	-1.28%	-4.93%	-4.92%	-2.66%	-6.98%	-6.95%
Class C	-2.13%	-4.76%	-4.85%	-1.89%	-6.34%	-6.62%	-4.97%	-11.01%	-11.78%
Class D	-3.08%	-5.24%	-6.90%	-2.28%	-7.91%	-8.85%	-7.30%	-13.62%	-16.43%
Average	**-1.54%**	**-3.67%**	**-4.12%**	**-1.70%**	**-5.85%**	**-5.89%**	**-3.94%**	**-8.23%**	**-9.23%**

TABLE III
BD-RATE RESULTS OF ABLATION STUDIES

Class	BD-rate											
	No IFA			No SFA			No FF			Ours		
	Y	U	V	Y	U	V	Y	U	V	Y	U	V
Class B	-2.79%	-6.32%	-5.97%	-2.70%	-9.38%	-8.84%	-2.73%	-7.43%	-7.08%	-2.97%	-7.54%	-7.25%
Class C	-4.96%	-12.11%	-11.46%	-4.53%	-14.49%	-14.67%	-5.13%	-11.65%	-11.84%	-5.51%	-11.73%	-12.30%
Class D	-7.21%	-13.16%	-14.08%	-6.43%	-15.27%	-16.76%	-6.61%	-13.05%	-14.65%	-7.17%	-13.47%	-16.08%
Average	**-4.99%**	**-10.53%**	**-10.50%**	**-4.55%**	**-13.04%**	**-13.39%**	**-4.67%**	**-10.24%**	**-10.87%**	**-5.22%**	**-10.91%**	**-11.88%**

TABLE IV
PERFORMANCE AND COMPLEXITY COMPARISON WITH OTHER WORKS

	Jia [27]	Jia [29]	**Ours**
Average BD-Rate of Y	-1.70%	-4.69%	**-3.94%**
Number of Parameters (M)	44.7	7.8	**7.8**
Multiply Accumulate(kMAC/pixel)	2843.8	1048	**729.8**

plexity. For consistent comparison, the experimental results were obtained using the same four QP values (22, 27, 32, 37) as previously mentioned. It should be noted that study [29], published in TCSVT, is the following study of [27] and represented the state-of-the-art model at the time of its publication.

The experimental results demonstrate that our method exhibits significant advantages over [27] in terms of both parameter count and computational complexity. When compared with [29], our approach achieves a 30% reduction in computational complexity while maintaining comparable coding efficiency, with only a 16% performance degradation in the Y component (and even smaller losses for U/V components). This substantial reduction in computational requirements facilitates practical deployment scenarios where maintaining coding efficiency is paramount.

D. Ablation Study

We systematically evaluate the contributions of three core modules through ablation studies. The study is performed on VTM-15.0 platform with five QPs as in Section III-B, and Class B, C and D are used as test sequences. Results are shown in Table III. The removal of the feature filtering module (No FF) results in an average BD-rate degradation of 0.55% in Y component, demonstrating its role in suppressing compression artifacts while preserving essential texture features. Disabling the inter-frame attention module (No IFA) leads to weakened temporal correlation modeling capability, causing a degradation of 0.23%. While the absence of the spatial-frequency attention (No SFA) slightly improves U/V component, it causes critical deterioration in Y as 0.67%. The result confirms its effectiveness in cross-component trade-off through frequency-domain attention mechanism, strengthening luma component which is more sensitive to human visual system. Experimental results collectively demonstrate that these three modules synergistically enhance coding efficiency.

IV. CONCLUSION

This work presents a deep learning-based approach that effectively enhances bidirectional inter-frame coding in H.266/VVC. By developing inter-frame attention mechanism and spatial-frequency processing modules, we address critical limitations in current DRF generation methods regarding high-level feature utilization and frequency-domain optimization. Experimental results on VTM-15.0 show 4.20%/8.53%/9.34% BD-rate reduction for Y/U/V components over the anchor which validate the framework's effectiveness, demonstrating significant BD-rate improvements while maintaining computational efficiency compared to existing works. The proposed solution offers a practical advancement in neural network-based video coding, with potential applications in next-generation video compression standards.

REFERENCES

[1] Thomas Wiegand, Gary J Sullivan, Gisle Bjøntegaard, and Ajay 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, 2003.

[2] Gary J Sullivan, Jens-Rainer Ohm, Woo-Jin Han, and Thomas 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, 2012.

[3] Benjamin Bross, Ye-Kui Wang, Yan Ye, Shan Liu, Jianle Chen, Gary J Sullivan, and Jens-Rainer Ohm, "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, 2021.

[4] Guo Lu, Wanli Ouyang, Dong Xu, Xiaoyun Zhang, Chunlei Cai, and Zhiyong Gao, "Dvc: An end-to-end deep video compression framework," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 11006–11015.

[5] Jiahao Li, Bin Li, and Yan Lu, "Deep contextual video compression," *Advances in Neural Information Processing Systems*, vol. 34, pp. 18114–18125, 2021.

[6] Shuai Huo, Dong Liu, Feng Wu, and Houqiang Li, "Convolutional neural network-based motion compensation refinement for video coding," in *IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2018, pp. 1–4.

[7] Zhenghui Zhao, Shiqi Wang, Shanshe Wang, Xinfeng Zhang, Siwei Ma, and Jiansheng 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, 2018.

[8] Yang Wang, Xiaopeng Fan, Ruiqin Xiong, Debin Zhao, and Wen Gao, "Neural network-based enhancement to inter prediction for video coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 2, pp. 826–838, 2021.

[9] Philipp Merkle, Martin Winken, Jonathan Pfaff, Heiko Schwarz, Detlev Marpe, and Thomas Wiegand, "Spatio-temporal convolutional neural network for enhanced inter prediction in video coding," *IEEE Transactions on Image Processing*, 2024.

[10] Han Zhang, Li Song, Zhengyi Luo, and Xiaokang Yang, "Learning a convolutional neural network for fractional interpolation in hevc inter coding," in *IEEE Visual Communications and Image Processing (VCIP)*. IEEE, 2017, pp. 1–4.

[11] Ning Yan, Dong Liu, Houqiang Li, Bin Li, Li Li, and Feng Wu, "Convolutional neural network-based fractional-pixel motion compensation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 3, pp. 840–853, 2018.

[12] Simon Niklaus, Long Mai, and Feng Liu, "Video frame interpolation via adaptive separable convolution," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 261–270.

[13] Jiaying Liu, Sifeng Xia, and Wenhan Yang, "Deep reference generation with multi-domain hierarchical constraints for inter prediction," *IEEE Transactions on Multimedia*, vol. 22, no. 10, pp. 2497–2510, 2019.

[14] Lei Zhao, Shiqi Wang, Xinfeng Zhang, Shanshe Wang, Siwei Ma, and Wen Gao, "Enhanced motion-compensated video coding with deep virtual reference frame generation," *IEEE Transactions on Image Processing*, vol. 28, no. 10, pp. 4832–4844, 2019.

[15] Sifeng Xia, Wenhan Yang, Yueyu Hu, and Jiaying Liu, "Deep inter prediction via pixel-wise motion oriented reference generation," in *IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 1710–1774.

[16] Hyomin Choi and Ivan V Bajić, "Affine transformation-based deep frame prediction," *IEEE Transactions on Image Processing*, vol. 30, pp. 3321–3334, 2021.

[17] Ziwei Liu, Raymond A Yeh, Xiaoou Tang, Yiming Liu, and Aseem Agarwala, "Video frame synthesis using deep voxel flow," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 4463–4471.

[18] Huaizu Jiang, Deqing Sun, Varun Jampani, Ming-Hsuan Yang, Erik Learned-Miller, and Jan Kautz, "Super slomo: High quality estimation of multiple intermediate frames for video interpolation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 9000–9008.

[19] Tarun Kalluri, Deepak Pathak, Manmohan Chandraker, and Du Tran, "Flavr: Flow-agnostic video representations for fast frame interpolation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2023, pp. 2071–2082.

[20] Lingtong Kong, Boyuan Jiang, Donghao Luo, Wenqing Chu, Xiaoming Huang, Ying Tai, Chengjie Wang, and Jie Yang, "Ifrnet: Intermediate feature refine network for efficient frame interpolation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 1969–1978.

[21] Zhewei Huang, Tianyuan Zhang, Wen Heng, Boxin Shi, and Shuchang Zhou, "Real-time intermediate flow estimation for video frame interpolation," in *European Conference on Computer Vision (ECCV)*. Springer, 2022, pp. 624–642.

[22] Zhen Li, Zuo-Liang Zhu, Ling-Hao Han, Qibin Hou, Chun-Le Guo, and Ming-Ming Cheng, "Amt: All-pairs multi-field transforms for efficient frame interpolation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 9801–9810.

[23] Guozhen Zhang, Yuhan Zhu, Haonan Wang, Youxin Chen, Gangshan Wu, and Limin Wang, "Extracting motion and appearance via inter-frame attention for efficient video frame interpolation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 5682–5692.

[24] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman, "Video enhancement with task-oriented flow," *International Journal of Computer Vision*, vol. 127, pp. 1106–1125, 2019.

[25] Wenbo Bao, Wei-Sheng Lai, Xiaoyun Zhang, Zhiyong Gao, and Ming-Hsuan Yang, "Memc-net: Motion estimation and motion compensation driven neural network for video interpolation and enhancement," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 3, pp. 933–948, 2019.

[26] Yuzhang Hu, Wenhan Yang, Jiaying Liu, and Zongming Guo, "Deep inter prediction with error-corrected auto-regressive network for video coding," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 19, no. 1s, pp. 1–22, 2023.

[27] Jianghao Jia, Zizheng Liu, Xiaozhong Xu, Shan Liu, and Zhenzhong Chen, "Deep reference frame interpolation based inter prediction enhancement for versatile video coding," in *IEEE International Conference on Visual Communications and Image Processing (VCIP)*. IEEE, 2022, pp. 1–5.

[28] Weijie Bao, Jianghao Jia, Wenhui Meng, Zizheng Liu, Xiaozhong Xu, Shan Liu, and Zhenzhong Chen, "Towards deep reference frame in versatile video coding nnvc," in *IEEE International Conference on Visual Communications and Image Processing (VCIP)*. IEEE, 2023, pp. 1–5.

[29] Jianghao Jia, Yuantong Zhang, Han Zhu, Zhenzhong Chen, Zizheng Liu, Xiaozhong Xu, and Shan Liu, "Deep reference frame generation method for vvc inter prediction enhancement," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 5, pp. 3111–3124, 2023.

[30] Wenhui Meng, Yuantong Zhang, Jianghao Jia, Songtao Chao, and Zhenzhong Chen, "Towards lightweight deep reference frame for versatile video coding," in *IEEE International Conference on Visual Communications and Image Processing (VCIP)*. IEEE, 2023, pp. 1–5.

[31] Chengzhuo Gui, Yuantong Zhang, Weijie Bao, Zhenzhong Chen, Huairui Wang, and Shan Liu, "Deep reference frame for versatile video coding with structural re-parameterization," in *IEEE International Conference on Visual Communications and Image Processing (VCIP)*. IEEE, 2024, pp. 1–5.

[32] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang, "Image super-resolution using deep convolutional networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, pp. 295–307, 2015.

[33] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-assisted Intervention (MICCAI)*. Springer, 2015, pp. 234–241.

[34] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu, "Eca-net: Efficient channel attention for deep convolutional neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 11534–11542.

[35] Zequn Qin, Pengyi Zhang, Fei Wu, and Xi Li, "Fcanet: Frequency channel attention networks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 783–792.

[36] Elena Alshina, Ru-Ling Liao, Shan Liu, and Andrew Segall, "Jvet common test conditions and evaluation procedures for neural network-based video coding technology," in *JVET 24th Meeting, document JVET-AB2016*, 2022, pp. 20–28.