An Extended Skip Strategy for Inter Prediction

Kuang Zhuo
Huazhong University of Science and Technology
2 PUBLICATIONS 0 CITATIONS
SEE PROFILE

Hongkui Wang
Huazhong University of Science and Technology
16 PUBLICATIONS 8 CITATIONS
SEE PROFILE

Xiaofeng Huang
Hangzhou Dianzi University
22 PUBLICATIONS 34 CITATIONS
SEE PROFILE

An Extended Skip Strategy for Inter Prediction

Hao Tao[1], Li Yu[1*], Zhuo Kuang[1], Hongkui Wang[1], Xiaofeng Huang[2]
[1]*School of Electron. Inf. & Commun., Huazhong Univ. of Sci. & Tech. Wuhan, Hubei, China*
[2]*School of Communication Engineering, Hangzhou Dianzi University, Hangzhou, China*
{husthtao, hustlyu, kuangzhuo, hkwang}@hust.edu.cn[1], xfhuang@hdu.edu.cn[2]

***Abstract*—The High Efficiency Video Coding (HEVC) standard adopts inter prediction to eliminate temporal correlation between the successive frames. However, a large amount of bits need to be explicitly signaled in the bitstream to specify the motion information. In this paper, we propose an extended skip strategy to alleviate bit consumption for motion data during the inter prediction process. Specifically, before the current frame is encoded, an additional picture generated by a deep convolutional neural network (CNN) is introduced to inter prediction. Since the additional reference picture is more similar with the current frame, most blocks of this frame can be skipped in the coding process. Consequently, to further improve the compression, an extended skip strategy is designed, i.e., the current frame can be skipped in multi-levels, including frame-level and coding tree unit level (CTU-level). Moreover, the skip-level of the current frame is decided in the sense of rate-distortion optimization (RDO). The proposed algorithm is implemented on the HM-16.6 software and an average of 4.4% BD-rate gain has been achieved in the experiments, which indicates the superiority of the proposed method.**

***Index Terms*—inter prediction, convolutional neural network, video coding**

I. Introduction

The most popular video coding standard, High Efficiency Video Coding (HEVC) [1], has adopted the block based hybrid coding framework. Inter prediction, which is aimed at eliminating temporal redundancy between consecutive frames, is one of the most crucial modules in HEVC. Specifically, inter prediction is implemented by performing motion search on the reference pictures to generate an optimal prediction block for the current block. Generally, higher prediction accuracy will lead to better coding performance, whereas more accurate prediction will consume more bits for motion information.

In view of this, extensive researches are devoted to improving the inter prediction accuracy to achieve better coding performance. Multiple works [2]–[6] attempted to utilize higher order motion models to more accurately compensate for the current block. In [2]–[4], various higher order motion models have been utilized to estimate the complex motions in the videos for better prediction accuracy. Although they could provide prediction blocks of higher quality, extra bits need to be allocated to higher order motion parameters. The other works preferred to perform motion compensation on a smaller block to obtain the optimal prediction performance. Lin *et al.* [5] proposed a paradigm of adaptive interpolated motion compensation (AIMC) to accurately compensate for off-grid blocks. Li *et al.* [6] utilized optical flow estimation to build a pixel-wise motion filed for better compensation performance. Prediction accuracy can be significantly improved by the aforementioned algorithms. Yet, more bits will be consumed for motion information, which may degrade the overall coding performance to a certain extent.

In order to decrease bit consumption for motion information, extensive researches have attempted to derive motion information at less bit cost. Li *et al.* [7] proposed a Motion Vector (MV) derivation algorithm based on frame rate up-conversion (FRUC) and no MV data need to be coded when FRUC mode is enabled. Yet, the MV derivation is not reliable in scenarios with complicated motions. Wu *et al.* [8] designed an incredible end-to-end deep learning codec which relied on repeated image interpolation and the performance of the codec is comparable to AVC [9]. Choi *et al.* [10] trained a deep neural network (DNN) model to generate a deep frame for inter prediction. When the deep frame is enabled, only the deep frame prediction flag and the residual data need to be coded. Lei *et al.* [11] introduced the adaptive separable convolution algorithm to generate a virtual reference frame to compensate for the current block. The algorithm mainly focused on CTU-level coding and could encode a block without bit consumption for motion information. Though the previous works have made great progress in alleviating bit consumption for motion data, it still can be further improved.

In this paper, in order to further decrease bit cost for motion data, a novel extended skip strategy is proposed for inter prediction, which leads to significant coding performance gain. Specifically, we first employ a deep CNN-based video frame interpolation algorithm to generate a high quality additional reference picture for better prediction. Since the prediction block obtained from the additional reference picture is of high quality, most blocks in the current frame can be skipped in the coding process. Consequently, a novel multi-level skip decision scheme is devised to accurately decide whether skipping the current block or frame. More specifically, a frame-level and CTU-level combined RDO strategy is proposed to adaptively make the skip decision. In addition, a novel coding mode, termed as extended skip (ES) mode, is correspondingly proposed to serve as a candidate inter prediction mode.

The remainder of this paper is organized as follows: A brief review of video frame interpolation algorithm will be presented in Section II. The proposed extended skip strategy will be concretely described in Section III. Experimental results will be given and analyzed in Section IV. Finally, Section V will conclude this paper.

*This work was supported by the NSFC (NO. 61871437)

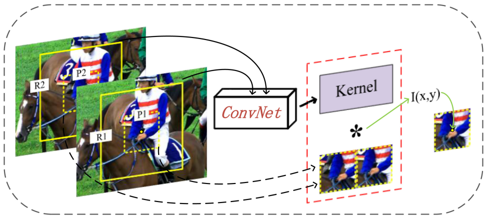


Fig. 1. Illustration of the deep CNN-based video frame interpolation method.

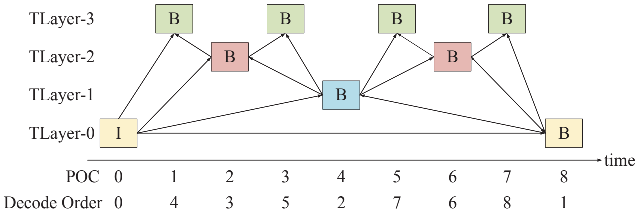


Fig. 2. Illustration of hierarchical B frame coding structure.

II. Review of Video Frame Interpolation Algorithm

Video frame interpolation is a classical task in video processing. Traditional video frame interpolation is a two-step process which includes dense motion estimation and pixel synthesis [12]–[14]. The interpolation performance is seriously dependent on the accuracy of any one of the two steps. Nevertheless, it is a huge challenge to estimate optical flow in scenarios with occlusion, blur and abrupt brightness change. Moreover, it is impractical to synthesize the target pixel accurately when occlusion happens.

With the rapid development of deep learning, video frame interpolation algorithm has made great progress. Extensive researches have demonstrated the superiority of applying deep learning to video frame interpolation and the deep CNN-based single step video frame interpolation method is an outstanding branch among them. Niklaus *et al.* [15] found that video frame interpolation process is essentially identical with the local convolution process. Therefore, they proposed to utilize an adaptive convolution algorithm to directly generate the intermediate frame, which combined the aforementioned two steps into a single convolution process. More specifically, the principle of the one-step method is shown in Fig. 1. In order to obtain the pixel (x, y) in the intermediate picture, two receptive field patches R1 and R2 centered at (x, y) are fed into the CNN to estimate a convolution kernel, which then convolves with the input patches P1 and P2 to synthesize the target pixel. Nevertheless, the deep CNN-based video frame interpolation method is not suitable for real-time applications due to its high computational complexity. Consequently, Niklaus *et al.* [16] improved their algorithm with an adaptive separable convolution algorithm which utilized a pair of 1-D kernels to approximate the 2-D kernel. In particular, a n×n convolution kernel can be encoded using only 2n variables, which is far less than the computational consumption of the 2-D kernel.

In view of the natural similarity between video frame interpolation and bi-directional B frame prediction, we propose to introduce the video frame interpolation algorithm based on adaptive separable convolution to inter prediction. In this manner, a high quality additional reference picture will be generated for better prediction performance.

III. The Proposed Method

Skip mode is a special merge mode in HEVC. When it is enabled, only merge index and MV index need to be signaled in the bitstream. However, the skip mode is only enabled when the current block is identical with its reference block. The skip condition is so rigorous that only blocks with little texture can meet it. In order to skip more blocks for improving the coding performance, an extended skip strategy is proposed in this section. More specifically, we propose to generate the additional reference picture F_{ARP} from two reconstructed reference pictures for better prediction performance. Since the quality of the addition reference picture is comparable with that of reconstruction picture generated by HEVC, it is practical to skip more blocks. Consequently, an extended skip strategy which can be performed in frame-level and CTU-level is proposed to adaptively decide whether skipping the frame or block in the sense of RDO. In addition, a novel coding mode, namely ES mode, is proposed correspondingly. More details will be presented in the rest of the section.

A. Additional Reference Picture Generation

As mentioned above, we propose to perform a video frame interpolation algorithm based on Adaptive Separable Convolution to generate the additional reference picture F_{ARP}. In particular, the F_{ARP} generation method is mainly applied in the hierarchical B frame coding structure [17], for both video frame interpolation and hierarchical B frame prediction are aimed at generating an intermediate frame. As shown in Fig. 2, a typical hierarchical B frame coding structure consists of four temporal layers. Frames in lower layers serve as reference pictures for frames in higher layers.

Once frames in lower levels have been reconstructed, the reconstructed reference pictures will be fed into the deep CNN to generate an additional reference picture for the current frame. As shown in Fig. 3 (a), assuming that the current frame F_x is a TLayer-3 frame and its nearest reference frames $F_{x+\Delta x}$ and $F_{x-\Delta x}$ have been reconstructed, the additional reference picture can be directly generated with the deep CNN-based video frame interpolation algorithm. Obviously, the additional reference pictures of the rest of TLayer-3 frames can be effectively generated in the same way. Let G(·) denote the additional reference picture generation process and Rec(·) denote the frame reconstruction process, the proposed generation process can be formulated as follow:

$$\widetilde{F_x} = G(Rec(F_{x-\Delta x}), Rec(F_{x+\Delta x})) \quad (1)$$

where $\widetilde{F_x}$ denotes the additional reference picture for the current frame F_x. The subscript x and Δx represent the POC of the current frame and the temporal distance between current frame and its reference, respectively. Obviously, the value of

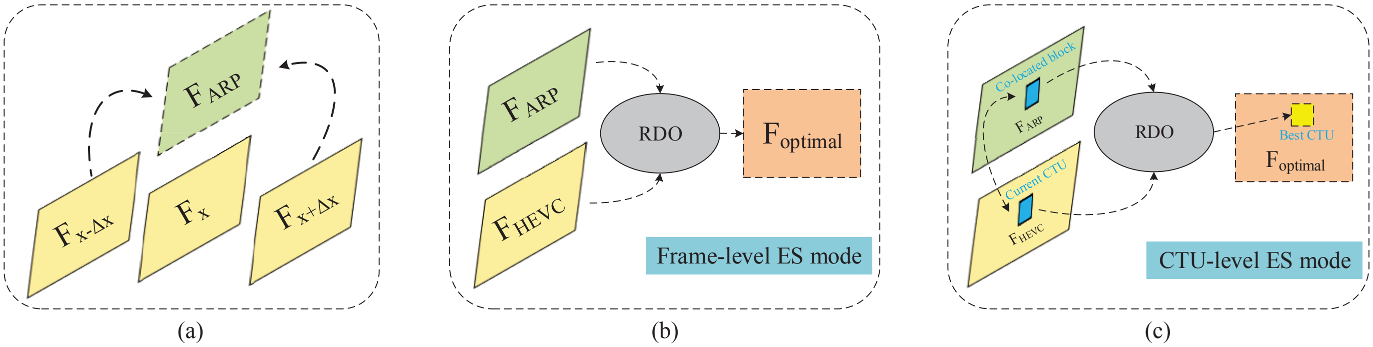


Fig. 3. Illustration of the proposed method. (a) F_{ARP} Generation: generate an additional reference picture by the deep CNN-based video interpolation algorithm. (b) Frame-level ES mode: check if F_{ARP} is the optimal prediction for the current frame. If so, skip the current frame. (c) CTU-level ES mode: check if the co-located block on F_{ARP} is the optimal prediction for the current block. If so, skip the current block.

Δx is getting smaller when the temporal layer level getting higher.

Once an additional reference picture has been generated, there is an urgent demand to design a strategy to fully exploit it. An intuitive scheme is to directly take the additional reference picture as the optimal prediction picture. In this manner, we can skip the frame in the coding process. If it fails to skip the frame, we will attempt to obtain better prediction blocks from the additional reference picture. Similarly, we can skip the blocks which obtained the optimal prediction blocks from the additional reference picture. Consequently, an extended skip strategy is devised to adaptively make skip decision in the sense of RDO, which will be introduced in detail in next subsection.

B. The Extended Skip Strategy

For the purpose of fully exploiting the additional reference picture, we firstly propose a novel coding mode, termed as extended skip (ES) mode, which serves as a candidate prediction mode. Furthermore, a frame-level and CTU-level combined skip strategy is correspondingly devised to decide whether the proposed ES mode is enabled in the sense of RDO. In contrast to classical inter prediction methods in HEVC, only a ES flag is required to be signaled in the bitstream when the ES mode is enabled, which brings considerable bitrate saving.

1) ES mode: The proposed ES mode serves as a candidate prediction mode when the current frame can be bi-predicted. After all normal prediction modes have been checked during the RDO process, the ES mode will be performed continuously. In particular, only when the temporal distances between the current frame and its reference pictures are identical could the ES mode be performed. As shown in Fig. 3 (b) and Fig. 3 (c), the proposed ES mode is carried out in frame-level and CTU-level, respectively. Once a high quality additional reference picture F_{ARP} has been generated, the ES mode is orderly performed in frame-level and CTU-level. More specifically, if the frame-level ES mode is chosen as the optimal mode, the F_{ARP} is directly employed as the optimal prediction frame and the current frame can be skipped. Otherwise, the CTU-level ES mode will be sequentially conducted to further achieve the optimal coding performance.

2) Frame-level and CTU-level Combined Skip Strategy: In order to determine whether the proposed ES mode is enabled, a frame-level and CTU-level combined RDO strategy is devised to make skip decision. The rate-distortion (RD) cost in HEVC can be formulated as follow:

$$J_{HEVC} = D_{HEVC} + \lambda \times R_{HEVC} \tag{2}$$

where J_{HEVC}, D_{HEVC} and R_{HEVC} denote the RD cost, distortion and bitrate of normal prediction mode in HEVC, respectively. The parameter λ represents Lagrange constant, which is utilized to control the relationship of bitrate and distortion. Similarly, the RD cost of the proposed ES mode is denoted as J_{ES} and it can be formulated as follow:

$$J_{ES} = D_{ES} + \lambda \times R_{ES} \tag{3}$$

where D_{ES} and R_{ES} denote the distortion and bitrate of ES mode, respectively. Since there is no bit consumption for motion information when ES mode is enabled, the bit expense of ES flags accounts for the majority of the overall bitrate of the current block.

The flowchart of the proposed skip strategy is shown in Fig. 4. Immediately after the high quality additional reference picture F_{ARP} is obtained, the frame-level RDO process is performed by comparing the J_{ES} and J_{HEVC} of current frame. When $J_{ES} < J_{HEVC}$ in frame-level, the frame-level ES mode is chosen as the optimal coding mode of the current frame and the frame-level ES flag is set to be true correspondingly. Otherwise, the CTU-level RDO process will be conducted continuously. Specifically, we firstly obtain the co-located block of the current CTU in F_{ARP} and calculate its RD cost. Afterwards, if $J_{ES} < J_{HEVC}$ in CTU-level, the CTU-level ES mode is selected as the optimal prediction mode and the CTU-level ES flag is set to be true correspondingly. In addition, traditional inter prediction mode will be chosen as the optimal coding mode when ES mode is suboptimal in both frame-level and CTU-level.

It is worth noting that the motion data and residual data do not need to be signaled in the bitstream when the ES mode is selected. In this manner, considerable bitrate saving can be achieved, which makes ES mode a competitive one among all candidate prediction modes in HEVC.

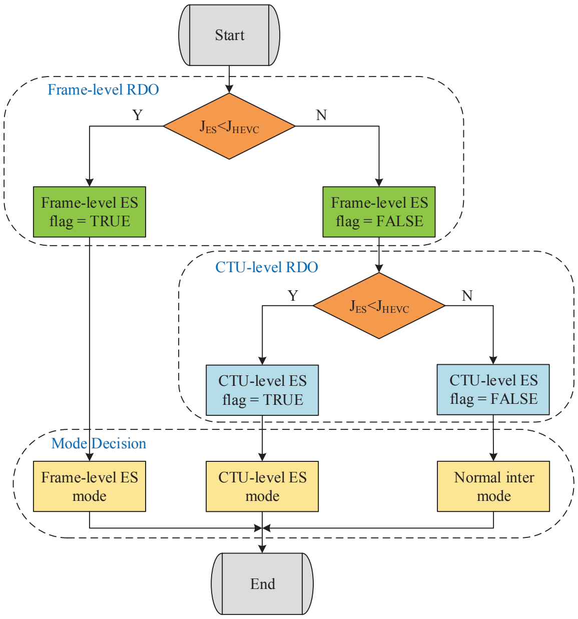


Fig. 4. Flowchart of the proposed skip strategy.

IV. Experimental Results

In order to verify the proposed algorithm, the Adaptive Separable Convolution [18] model is integrated into HM-16.6 to generate an additional reference picture for ES mode. In order to compare with methods in [11], the tested quantization parameters are set to be 27, 32, 37 and 42. Each sequence is encoded for 2 seconds with the Random Access (RA) configuration and only luma component is evaluated in the experiments. In addition, other test conditions are consistent with the Common Test Conditions [19].

As is introduced in section III, we can generate an additional reference picture of higher quality with the deep CNN-based video frame interpolation algorithm when the temporal distance between the input frames getting closer. Consequently, the proposed ES mode is only applied to TLayer-3 frames in our first experiment. Nevertheless, any B frame which has bi-directional reference pictures can employ the ES mode as a candidate prediction mode. Generally speaking, it is not preferred to applying the ES mode to B frames in lower layers, for they will be referenced by frames in higher layers. In order to evaluate the coding performance of applying the ES mode to B frames in lower temporal layers, the ES mode is applied to both TLayer-2 and TLayer-3 frames in another experiment. Moreover, the coding performance is evaluated with Bjontegaard-Delta rate (BD-rate) [20] in all experiments.

As shown in Table I, the proposed method achieves an average of 2.7% BD-rate saving when only applying ES mode to TLayer-3 frames. The BD-rate saving on BQSquare is up to 5.6%, which demonstrates the superiority of the proposed method. Compared with coding performance of Zhao *et al.* [11], our method provides the best performance in 8 out of 13 sequences and achieves an average of 0.4% BD-rate gain. However, the coding performance on RaceHorsesC is only slightly better than HEVC, which indicates that the deep CNN-based interpolation algorithm can not generate a high quality additional reference picture for videos with complex motion.

As we can see from Table II, the coding performance has achieved further improvement against the first experiment. More specifically, an average of 4.4% BD-rate saving is achieved, which verifies the effectiveness of applying ES mode to TLayer-2 frames. It is incredible that up to 10.7% BD-rate saving is achieved on BQSquare, which provides 3.6% BD-rate gain compared with that of Zhao *et al.* [11]. In comparison with method in [11], our method achieves best performance in 10 of 13 sequences and an average of 1.2% BD-rate gain can be observed. Since Zhao *et al.* [11] only apply DVRF mode in CTU-level while our method can perform ES mode in frame-level, hundreds of bits per frame can be saved when frame-level ES mode is enabled. Besides, the reference quality becomes worse in HEVC when the temporal distances between the current frame and its references getting farther, which leads to more bit consumption for TLayer-2 frames. However, the deep CNN-based interpolation algorithm still can generate high quality reference pictures for TLayer-2 frames. Consequently, significant coding performance improvement can be achieved in this manner.

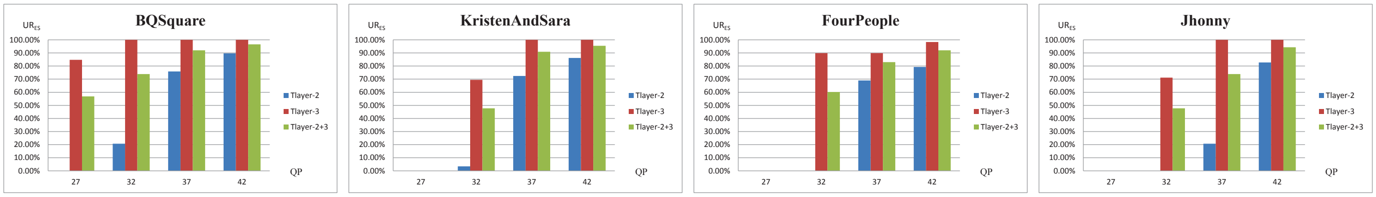
In order to further explore the performance difference on the test sequences, we calculate the utilization rate of ES mode in different sequences. Let UR_{ES} denote the utilization rate of ES mode, it can be formulated as follow:

$$UR_{ES} = \frac{N_{ES}}{N_{ALL}} \times 100\% \tag{4}$$

where N_{ES} and N_{ALL} denote frame-level ES mode times and total number of frames that enable ES mode, respectively. As shown in Fig. 5, we can observe that the UR_{ES} varies among the test sequences and increases as QP gets larger. More specifically, the UR_{ES} of BQSquare is rather high in all QPs, which leads to the best coding performance. Regarding KristenAndSara, FourPeople and Johnny, almost no frames select frame-level ES mode as the optimal coding mode in low QP setting while over 70% frames select frame-level ES mode when QP getting larger, which indicates that the deep CNN-based interpolation algorithm can not generate a high quality additional reference picture in high quality videos.

V. Conclusion

In this paper, we present an extended skip strategy for inter prediction to improve the coding performance in HEVC. It is remarkable that an additional reference picture is introduced to inter prediction. Since the additional reference picture is more similar with the current frame, most blocks of this frame can be skipped in the coding process. Consequently, a multi-level extended skip strategy is devised to adaptively make skip decision in the sense of RDO. In this manner, better coding performance can be achieved without bit consumption for motion information. A significant BD-rate saving has been observed in the experiments, which shows that the proposed method outperforms the existing methods.



(a) (b) (c) (d)

Fig. 5. The utilization rate of frame-level ES mode in different sequences.

TABLE I
CODING PERFORMANCE WHEN ONLY APPYING THE PROPOSED METHOD TO TLAYER-3 FRAMES.

Sequence		**BD-rate**	
		Proposed	***Zhao et al. [11]***
Class A	PeopleOnStreet	-3.3%	-
	Traffic	-2.1%	-
Class B	Kimono	-1.6%	-1.6%
	ParkScene	**-2.9%**	-2.2%
	Cactus	-2.5%	**-2.7%**
	BasketballDrive	**-1.1%**	-0.6%
	BQTerrace	**-0.8%**	+0.2%
Class C	BasketballDrill	-2.5%	**-2.6%**
	BQMall	**-4.8%**	-4.4%
	PartyScene	-2.4%	-2.4%
	RaceHorseC	**-0.6%**	-0.5%
Class D	BasketballPass	**-4.0%**	-3.9%
	BQSquare	**-5.6%**	-5.4%
	BlowingBubbles	-2.1%	**-3.0%**
	RaceHorse416	**-2.1%**	-1.6%
Class E	FourPeople	-4.7%	-
	Johnny	-2.1%	-
	KristenAndSara	-2.5%	-
Average		**-2.7%**	-2.3%

TABLE II
CODING PERFORMANCE WHEN APPYING THE PROPOSED METHOD TO TLAYER-2 AND TLAYER-3 FRAMES.

Sequence		**BD-rate**	
		Proposed	***Zhao et al. [11]***
Class A	PeopleOnStreet	-3.7%	-
	Traffic	-4.5%	-
Class B	Kimono	**-1.9%**	-1.7%
	ParkScene	**-5.2%**	-2.6%
	Cactus	**-4.9%**	-4.6%
	BasketballDrive	**-1.3%**	-1.1%
	BQTerrace	**-1.0%**	-0.2%
Class C	BasketballDrill	-2.9%	**-3.2%**
	BQMall	**-6.9%**	-6.0%
	PartyScene	**-3.7%**	-3.0%
	RaceHorseC	-0.8%	-0.8%
Class D	BasketballPass	-5.4%	-5.4%
	BQSquare	**-10.7%**	-7.1%
	BlowingBubbles	**-4.2%**	-4.1%
	RaceHorse416	**-2.4%**	-2.2%
Class E	FourPeople	-7.5%	-
	Johnny	-6.1%	-
	KristenAndSara	-6.6%	-
Average		**-4.4%**	-3.2%

REFERENCES

[1] 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, 2012.

[2] L. Li, H. Li, Z. Lv, and H. Yang, "An affine motion compensation framework for high efficiency video coding," in *International Symposium on Circuits and Systems*. IEEE, 2015, pp. 525–528.

[3] C. Heithausen and J. H. Vorwerk, "Motion compensation with higher order motion models for hevc," in *International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2015, pp. 1438–1442.

[4] C. Heithausen, M. Bläser, M. Wien, and J.-R. Ohm, "Improved higher order motion compensation in HEVC with block-to-block translational shift compensation," in *International Conference on Image Processing*. IEEE, 2016, pp. 2008–2012.

[5] W.-T. Lin, T. Nanjundaswamy, and K. Rose, "Adaptive interpolated motion-compensated prediction with variable block partitioning," in *Data Compression Conference*. IEEE, 2018, pp. 23–31.

[6] B. Li, J. Han, and Y. Xu, "Co-located reference frame interpolation using optical flow estimation for video compression," in *Data Compression Conference*. IEEE, 2018, pp. 13–22.

[7] X. Li, J. Chen, and M. Karczewicz, "Frame rate up-conversion based motion vector derivation for hybrid video coding," in *Data Compression Conference*. IEEE, 2017, pp. 390–399.

[8] C. Wu, N. Singhal, and P. Krahenbuhl, "Video compression through image interpolation," in *European Conference on Computer Vision*. IEEE, 2018, pp. 416–431.

[9] 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, 2003.

[10] H. Choi and I. V. Bajic, "Deep frame prediction for video coding," *arXiv preprint arXiv:1901.00062*, 2018.

[11] L. Zhao, S. Wang, X. Zhang, S. Wang, S. Ma, and W. Gao, "Enhanced ctu-level inter prediction with deep frame rate up-conversion for high efficiency video coding," in *International Conference on Image Processing*. IEEE, 2018, pp. 206–210.

[12] Z. Yu, H. Li, Z. Wang, Z. Hu, and C. W. Chen, "Multi-level video frame interpolation: Exploiting the interaction among different levels," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 7, pp. 1235–1248, 2013.

[13] M. Werlberger, T. Pock, M. Unger, and H. Bischof, "Optical flow guided TV-L1 video interpolation and restoration," in *International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition*. Springer, 2011, pp. 273–286.

[14] S. Baker, D. Scharstein, J. Lewis, S. Roth, M. J. Black, and R. Szeliski, "A database and evaluation methodology for optical flow," *International Journal of Computer Vision*, vol. 92, no. 1, pp. 1–31, 2011.

[15] S. Niklaus, L. Mai, and F. Liu, "Video frame interpolation via adaptive convolution," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2017, pp. 670–679.

[16] ——, "Video frame interpolation via adaptive separable convolution," in *IEEE International Conference on Computer Vision*. IEEE, 2017, pp. 261–270.

[17] H. Schwarz, D. Marpe, and T. Wiegand, "Analysis of hierarchical B pictures and MCTF." in *International Conference on Multimedia and Expo*. Citeseer, 2006, pp. 1929–1932.

[18] Github, "sepconv," in *http://graphics.cs.pdx.edu/project/sepconv*.

[19] K. Suehring and X. Li, "Jvet common test conditions and software reference configurations," in *JVET-G1010*, 2017.

[20] G. Bjontegaard, "Calculation of average PSNR differences between RD-Curves," *in VCEG-M33*, 2001.