Collaborative Decoder-side Motion Vector Refinement for Video Coding

Jialin Li, Zhuoyuan Li, Yao Li, Li Li, Houqiang Li
University of Science and Technology of China (USTC), Hefei, Anhui 230027, China

***Abstract*—Motion compensation prediction (MCP) is a key technology to reduce the temporal redundancy in video coding. Recently, in order to improve its efficiency, the decoder-side MCP schemes are gradually adopted in advanced video coding standards, especially the decoder-side motion vector refinement (DMVR). In DMVR, the bilateral matching scheme is used to refine the motion vector (MV) obtained from merge-based inter mode at the sub-block level, which assumes that the motion vector difference (MVD) in the two reference directions of a bi-prediction block has the symmetric property. Although the bilateral assumption can effectively reduce complexity without any extra signal, the fixed searching rules limit the motion vector accuracy. To address this limitation, we propose a collaborative decoder-side motion vector refinement (C-DMVR) framework. In C-DMVR, the sub-block-based collaborative mechanism is introduced to optimize the distortion calculation (CDC) and bilateral-based searching strategy (CSS) to avoid inaccurate searching results, respectively. In the CDC, the receptive field of the distortion function is enlarged with the collaboration of additional spatial neighbor information to assist the accurate decision of motion vector candidates. In the CSS, the coarse-to-fine candidate list derivation scheme is introduced to construct the accurate searching path with the collaboration of neighbor sub-blocks. The proposed method is implemented into the AOMedia Video 2 reference software, AV2. Experimental results show that the proposed method achieves, on average, -0.17% and up to -0.60% BD-rate reduction compared to the AV2 anchor under the random access (RA) configuration, with a slight increase of time complexity on the encoding/decoding side.**

***Index Terms*—collaborative decoder-side motion vector refinement, searching path, distortion calculation, video coding, AV2.**

I. Introduction

In modern video codecs, block-based motion compensation prediction (MCP) is widely used to model the temporal correlation between the pictures of a video sequence [1]–[6]. In the recent development of advanced standards, the decoder-side MCP schemes are tentatively studied and gradually adopted in the standardization to promote the prediction efficiency, such as the decoder-side motion vector derivation (DMVD), and decoder-side motion vector refinement (DMVR).

In DMVD, [7]–[10] proposes that the motion information can be derived using template matching-based algorithms at the decoder side, instead of explicitly transmitting motion information in the encoder. To reduce the computational complexity of its decoder-side searching, the bi-directional template matching-based schemes are investigated to optimize its searching strategy [11]. Although these methods can enable the practical application of DMVD to some extent, the high demand of decoder-side time complexity and extra bits consumption are unavoidable.

Based on the DMVD, DMVR is gradually proposed to be applied adaptively on top of the blocks predicted with merge-based inter mode without extra flags, which saves additional bitrate [12]–[14]. Instead of the template of neighboring decoded samples surrounding the current block in DMVD, DMVR is based on the reconstructed samples retrieved by the initial motion vectors (MVs) derived by the merge-based inter mode in the reference pictures.

For the DMVR applied in AV2, as illustrated in Fig. 1 (a), bilateral matching is applied to refine the initial MVs (MV_0, MV_1) obtained from the merge-based inter mode (NEAR_NEARMV mode) and JOINT_NEWMV mode at the sub-block level. The process of bilateral matching mainly contains two parts, including distortion calculation (DC) and search strategy (SS). For DC, as illustrated in Fig. 1 (b), the sum of the absolute difference (SAD) of the block pair is calculated as the distortion cost of MV pair (MV_0', MV_1'), and the MV pair with the minimum SAD corresponds to the refined MVs. For SS, as illustrated in Fig. 1 (c), the candidate list is constructed with the symmetric property of motion vector candidates (ΔMV). The refined MVs (MV_0', MV_1') can be defined as $MV_0' = MV_0 + \Delta MV$ and $MV_1' = MV_1 - \Delta MV$ in different reference directions.

DMVR assumes that the coding block is divided into sub-blocks, and the refinement is applied to each sub-block independently to search the optimal ΔMV within the fixed searching path, which is constructed with the symmetric property. Although the bilateral assumption can effectively reduce complexity without any extra signal and provide robustness for the searching results, the fixed searching rules limit the motion vector accuracy. For the limitation of DC, only the fixed calculation range of block pairs is considered in the distortion function for each sub-block, which leads to a limited receptive field. The limited receptive field may result in an inaccurate decision of ΔMV for different motion situations (extreme/slight case), so the receptive field of the distortion function needs to be enlarged for different cases. For the limitation of SS, the motion situation varies from different sub-blocks, so the fixed searching path limits the motion vector accuracy and may result in inaccurate searching results.

Inspired by the above two points, in this paper, we propose a C-DMVR framework that introduces the sub-block-based collaborative mechanism to optimize the refinement for each sub-block. In C-DMVR, the collaborative distortion calcula-

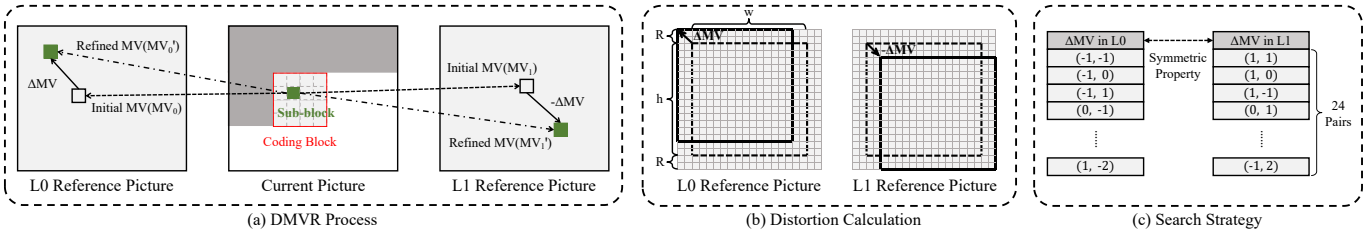

Fig. 1. Illustration of the framework of decoder-side motion vector refinement (*DMVR*). (a) The red block in the current picture is the coding block, the dotted blocks in the current picture are sub-blocks of the coding block and the green block in the current picture is the sub-block being processed. The gray block pair in reference pictures denotes the initial block pair retrieved by the initial MVs. The green block pair in reference pictures is the block pair with the minimum SAD cost compared with the other candidate block pairs or the initial block pair. (b) The dotted block pair and the solid block pair in the reference pictures denote the initial block pair and the candidate block pair, respectively. w and h are the width and height of the sub-block, and the search range is denoted by R. (c) The candidate list which is constructed with the symmetric property of ΔMV when $R = 2$.

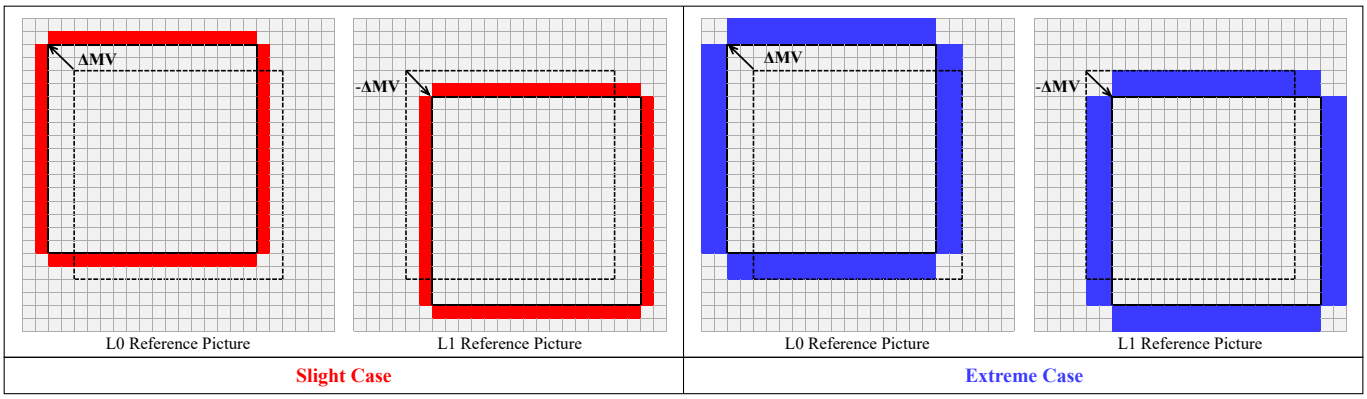

Fig. 2. Illustration of the collaborative distortion calculation (*CDC*). Different distortion functions are designed for different motion situations of each sub-block, including slight case and extreme case. For slight case, the red region introduced in the distortion function represents a 1-pixel width row or column. For extreme case, the blue region introduced in the distortion function represents a 2-pixel width row or column.

tion (CDC) and the collaborative searching strategy (CSS) are proposed to avoid inaccurate searching results. In CDC, the receptive field of the distortion function is enlarged with the collaboration of additional spatial neighbor information to assist the accurate decision of ΔMV. In CSS, the coarse-to-fine candidate list derivation scheme is introduced to construct the optimized candidate list with the collaboration of neighbor sub-blocks. Compared with the fixed searching path for all sub-blocks, the optimized candidate list considers different qualities of the potential ΔMVs to improve its accuracy.

The rest of the paper is organized as follows. In Section II, we introduce the proposed C-DMVR framework in detail. In Section III, the experimental results are shown and discussed. Section IV concludes this paper.

II. Proposed Method

In our method, the collaboration of neighboring spatial information is introduced in DMVR to improve the accuracy of the searching results in two aspects. For DC, CDC is proposed to enlarge the receptive field of distortion function. For SS, CSS is proposed to optimize the fixed searching path with the coarse-to-fine candidate list derivation scheme for each sub-block. In the following, we introduce them in detail.

A. Collaborative Distortion Calculation (CDC)

In the previous DMVR, the SAD of the block pairs retrieved by the refined MVs in the reference pictures is used for DC. This distortion function aims to estimate the block pair with the least distortion, but only the fixed calculation range of block pairs is considered, which leads to a limited receptive field. It may result in an inaccurate decision of ΔMV for different motion situations (extreme/slight case). To address this limitation, we propose CDC with an adaptive receptive field. For the design of CDC, the receptive field is enlarged with the collaboration of additional spatial neighbor information, which introduces the extra spatial neighbor pixels surrounding the block pairs. Thus, the definition of CDC is expressed as:

$$SAD = SAD(C0, C1) + SAD(N0, N1) \quad (1)$$

where $C0$ and $C1$ are the block pairs in the reference pictures, $N0$ and $N1$ are the spatial neighbor pixels surrounding the block pairs in the reference pictures.

For the process of CDC, first, the range of spatial neighbor pixels is determined. We assume that the worse the bilateral matching, the more spatial neighbor information needs to be introduced in the CDC. The distortion cost of the initial block pair (retrieved by the initial MVs) is calculated as the initial cost SAD_0 and used to classify the motion situation of each sub-block into extreme case or slight case. The classification rule is defined as slight case when the SAD_0 is larger than $2\times w\times h$ and less than $4\times w\times h$, and extreme case when the SAD_0 is larger than $4\times w\times h$. For different cases, the different distortion functions are designed to introduce spatial neighbor pixels. As illustrated in Fig. 2, for slight case, only the

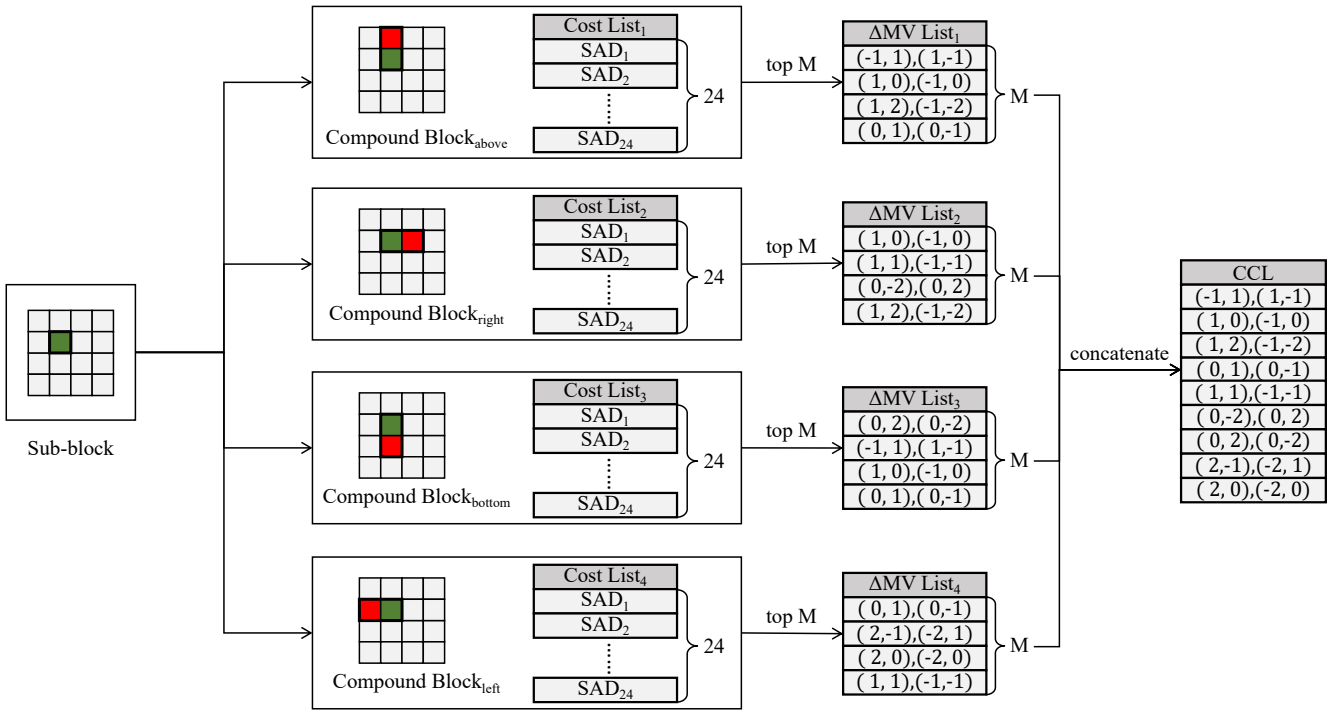

Fig. 3. Illustration of the collaborative search strategy (*CSS*) when search range $R = 2$. The sub-block being processed is colorized in green, and the neighbor sub-blocks used for collaboration are colorized in red. For each compound block, SAD_1-SAD_{24} in the cost list represents the distortion cost of 24 ΔMVs in the fixed searching path constructed with the symmetric property. For each compound block, the ΔMVs corresponding to the M minimum values in the cost list are selected to construct a pruned ΔMV list. Four pruned ΔMV lists can be obtained and concatenated to construct the CCL.

pixels within the 1-pixel width surrounding the block pairs are involved in the CDC. For extreme case, the pixels within the 2-pixel width surrounding the block pairs are involved in the CDC. Second, the above distortion function is used to calculate the distortion cost of ΔMVs for each sub-block.

B. Collaborative Search Strategy (CSS)

In the previous DMVR, the search range is defined as the number of integer pixels surrounding the initial block pair in both horizontal and vertical directions to limit the maximum value of ΔMV between the refined MVs and the initial MVs. For each sub-block, the candidate list is constructed with the symmetric property of ΔMV, and the number of candidate block pairs is equal to $(2R + 1)^2 - 1$, where R denotes the search range. Although the fixed searching path constructed with the symmetric property can effectively reduce the number of candidate block pairs and provide robustness for the searching results, the motion situation varies from different sub-blocks, so the fixed searching path limits the motion vector accuracy and may result in inaccurate searching results. To address this limitation, we propose CSS to optimize the fixed searching path at the sub-block level.

For the design of CSS, the coarse-to-fine candidate list derivation scheme is introduced to construct the collaborative candidate list (CCL), which considers different qualities of the potential ΔMVs with the collaboration of neighbor sub-blocks. The process of the construction of the CCL is shown in Fig. 3. It contains two stages, and the search range $R = 2$ case is shown as the example.

For the first stage, the candidate list is constructed with the symmetric property of ΔMV. For each sub-block, SAD values of the ΔMVs of the candidate list are calculated and stored in the SAD array for subsequent use. For the second stage, the current sub-block and its neighbor sub-block (one of the sub-blocks vertically or horizontally from the current sub-block) are combined and considered as a compound block. Four compound blocks and the corresponding cost list of the compound blocks can be derived from the SAD array generated in the first stage. For each compound block, the ΔMVs corresponding to the M minimum values in the cost list are selected to construct a pruned ΔMV list. Four pruned ΔMV lists can be obtained and concatenated to construct the CCL, which considers different qualities of the potential ΔMVs for the current sub-block with the collaboration of neighbor sub-blocks. Note that only the existing neighbor sub-block needs to be considered if the current sub-block is at the edge position of the current picture. Similar to the CDC, CSS is only applied to extreme motion cases.

III. Experimental Results

A. Experimental Settings

To evaluate the performance of C-DMVR, C-DMVR is implemented into the AOMedia Video 2 reference software, AV2 [15]. Since the DMVR is applied for bilateral matching with one reference picture in the past and one reference picture in the future, only the random access (RA) configuration is adopted as the test condition. The test sequences from classes B to E with different resolutions are tested as specified in

TABLE I
CODING PERFORMANCE AND RELATIVE COMPLEXITY OF C-DMVR BASED ON AV2 UNDER RA CONFIGURATION

Class	Sequence	Y	U	V
ClassB (1920x1080)	*MarketPlace*	0.03%	0.21%	0.20%
	RitualDance	-0.15%	-0.33%	-0.17%
	Cactus	-0.10%	-0.40%	-0.37%
	BasketballDrive	-0.21%	-0.64%	0.08%
	BQTerrace	-0.04%	-0.12%	-0.31%
ClassC (832x480)	*BasketballDrill*	-0.01%	-0.36%	0.69%
	BQMall	-0.33%	0.45%	-0.02%
	PartyScene	-0.17%	0.29%	-0.57%
	RaceHorses	-0.06%	-0.31%	0.49%
ClassD (416x240)	*BasketballPass*	-0.60%	-0.03%	0.64%
	BQSquare	0.03%	0.01%	-0.60%
	BlowingBubbles	-0.03%	-0.95%	1.07%
	RaceHorses	-0.35%	-0.14%	-0.43%
ClassE (1280x720)	*FourPeople*	-0.16%	-0.19%	0.23%
	Johnny	-0.33%	-0.63%	-0.32%
	KristenAndSara	-0.18%	-0.47%	-0.41%
Overall		**-0.17%**	**-0.23%**	**0.01%**
EncT		104%		
DecT		127%		

TABLE II
CODING PERFORMANCE AND RELATIVE COMPLEXITY OF CDC AND CSS BASED ON AV2 UNDER RA CONFIGURATION

Class	Sequence	CDC			CSS		
		Y	U	V	Y	U	V
ClassB (1920x1080)	*MarketPlace*	-0.07%	0.28%	0.25%	-0.01%	-0.07%	-0.03%
	RitualDance	-0.14%	-0.08%	0.01%	-0.16%	-0.15%	0.12%
	Cactus	0.04%	-0.10%	-0.68%	0.01%	0.19%	-0.35%
	BasketballDrive	-0.04%	-0.53%	0.00%	0.01%	0.18%	-0.30%
	BQTerrace	-0.03%	0.06%	-0.45%	-0.08%	-0.40%	-0.54%
ClassC (832x480)	*BasketballDrill*	-0.08%	-0.44%	0.16%	-0.02%	0.00%	0.78%
	BQMall	-0.11%	-0.90%	-0.35%	-0.10%	-0.06%	0.11%
	PartyScene	-0.23%	0.60%	-0.96%	-0.10%	-0.19%	-0.74%
	RaceHorses	-0.01%	-0.75%	-0.38%	0.03%	0.03%	0.32%
ClassD (416x240)	*BasketballPass*	-0.21%	0.50%	0.28%	-0.08%	0.36%	-0.48%
	BQSquare	-0.02%	-0.94%	0.19%	0.00%	-0.01%	-0.06%
	BlowingBubbles	-0.34%	-1.17%	0.31%	-0.04%	-1.01%	0.48%
	RaceHorses	-0.14%	-0.26%	-0.73%	-0.10%	0.24%	-0.26%
ClassE (1280x720)	*FourPeople*	0.05%	-0.34%	0.50%	-0.01%	-0.02%	0.42%
	Johnny	-0.40%	-0.62%	-0.59%	-0.24%	-0.53%	0.01%
	KristenAndSara	-0.08%	0.23%	-0.52%	0.01%	0.11%	-0.17%
Overall		**-0.11%**	**-0.28%**	**-0.19%**	**-0.06%**	**-0.08%**	**-0.04%**
EncT		99%			99%		
DecT		114%			104%		

[16]. Following the AV2 Common Test Condition (CTC) [17], quantization parameter (QP) values are set to 110, 135, 160, 185, 210, 235, and Bjontegaard Delta-rate (BD-rate) [18] is used as an objective metric to evaluate coding performance.

B. Performance

The R-D performance of the C-DMVR on the test sequences from classes B to E with different resolutions is illustrated in Table I. Y, U, and V represent the R-D performance gain of the three channels of YUV, EncT/DecT represent the encoding/decoding time. We can see that our proposed C-DMVR can achieve, on average, -0.17% and up to -0.60% BD-rate reduction (Y component) under the RA configuration. The experimental results show that the proposed framework performs better for sequences with drastic scene transformations and more moving objects, such as $BasketballPass$, $RaceHorses$, $BQMall$, and $BasketballDrive$.

For the time complexity, the encoding/decoding time ratio of the C-DMVR compared with the AV2 anchor is 104%/127%, under the RA configuration, respectively. The slight increase in EncT/DecT is mainly due to the introduced collaborative mechanism for the fine-grained candidate list derivation and cost distortion.

C. Ablation

To demonstrate the contributions of two core modules in C-DMVR, we conduct the ablation experiments on the proposed CDC and CSS under RA configuration.

First, we validate the effectiveness of the CDC, and only the CDC is turned on in the C-DMVR. As shown in Table II, we can see that our proposed CDC can achieve, on average, -0.11% BD-rate reduction (Y component) under the RA configuration. For the comparison of DC and CDC, the experimental results show that the CDC is superior to the DC, which demonstrates the collaboration of spatial neighbor information is beneficial to generate accurate decisions of ΔMV.

Second, we validate the effectiveness of CSS, and only the CSS is turned on in the C-DMVR. In Table II, we can see that our proposed CSS can achieve, on average, -0.06% BD-rate reduction (Y component) under the RA configuration. For the comparison of SS and CSS, the experimental results show that CSS is superior to the SS, which demonstrates the collaboration of neighbor sub-blocks is beneficial to construct a more accurate searching path.

IV. CONCLUSION

In this paper, the framework of C-DMVR is proposed to improve the existing DMVR technology in AV2. In C-DMVR, the CDC and CSS modules are introduced to optimize the two inherent steps (DC, SS) of existing DMVR. Experimental results show that the proposed C-DMVR can achieve, on average, -0.17% BD-rate reduction compared to the AV2 anchor with a slight increase of encoding/decoding time complexity under the RA configuration.

For future work, first, we will explore the more smart construction rule of collaborative candidate list construction to further assist in the accurate decoder-side bilateral matching. Second, we will extend the collaborative mechanism with more decoder-side inter prediction tools, such as the non-mirrored property-based DMVR [19], [20] and decoder-side affine model refinement (DAMR) [21], [22].

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] B. Bross, Y.-K. Wang, Y. Ye, S. Liu, J. Chen, G. J. Sullivan, and J.-R. 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.

[3] D. Mukherjee, J. Han, J. Bankoski, R. Bultje, A. Grange, J. Koleszar, P. Wilkins, and Y. Xu, "A technical overview of vp9—the latest open-source video codec," *SMPTE Motion Imaging Journal*, vol. 124, no. 1, pp. 44–54, 2015.

[4] J. Han, B. Li, D. Mukherjee, C.-H. Chiang, A. Grange, C. Chen, H. Su, S. Parker, S. Deng, U. Joshi *et al.*, "A technical overview of AV1," *Proceedings of the IEEE*, vol. 109, no. 9, pp. 1435–1462, 2021.

[5] Y. Li, Z. Li, L. Li, D. Liu, and H. Li, "Global homography motion compensation for versatile video coding," in *2022 IEEE International Conference on Visual Communications and Image Processing (VCIP)*. IEEE, 2022, pp. 1–5.

[6] Z. Li, Z. Yuan, L. Li, D. Liu, X. Tang, and F. Wu, "Object segmentation-assisted inter prediction for versatile video coding," *arXiv preprint arXiv:2403.11694*, 2024.

[7] S. Kamp, M. Evertz, and M. Wien, "Decoder side motion vector derivation for inter frame video coding," in *2008 15th IEEE International Conference on Image Processing*. IEEE, 2008, pp. 1120–1123.

[8] S. Kamp, B. Bross, and M. Wien, "Fast decoder side motion vector derivation for inter frame video coding," in *2009 Picture Coding Symposium*. IEEE, 2009, pp. 1–4.

[9] S. Kamp, J. Ballé, and M. Wien, "Multihypothesis prediction using decoder side-motion vector derivation in inter-frame video coding," in *Visual Communications and Image Processing (VCIP) 2009*, vol. 7257. SPIE, 2009, pp. 17–24.

[10] S. Kamp and M. Wien, "Decoder-side motion vector derivation for block-based video coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1732–1745, 2012.

[11] Y.-j. Chiu, L. Xu, W. Zhang, and H. Jiang, "Decoder-side motion estimation and wiener filter for HEVC," in *2013 Visual Communications and Image Processing (VCIP)*. IEEE, 2013, pp. 1–6.

[12] X. Chen *et al.*, "Decoder-side motion vector refinement based on bilateral template matching, jvet-d0029," in *4th Meeting: Chengdu, CN*, 2016.

[13] H. Gao, X. Chen, S. Esenlik, J. Chen, and E. Steinbach, "Decoder-side motion vector refinement in VVC: Algorithm and hardware implementation considerations," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 8, pp. 3197–3211, 2020.

[14] H. Gao, S. Esenlik, Z. Zhao, E. Steinbach, and J. Chen, "Decoder side motion vector refinement for versatile video coding," in *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*. IEEE, 2019, pp. 1–6.

[15] AOM, "AOMedia Video 1 and 2 (AV1 and AV2)," *https://aomedia.org/*, 2015.

[16] K. Suehring and X. Li, "JVET common test conditions and software reference configurations," *JVET-B1010*, 2016.

[17] X. Zhao, Z. Lei, A. Norkin, T. Daede, and A. Tourapis, "AOM common test conditions v2.0," *Alliance for Open Media, Codec Working Group Output Document*, 2021.

[18] G. Bjontegaard, "Calculation of average psnr differences between rd-curves," *ITU-T VCEG-M33, April, 2001*, 2001.

[19] Y. Jian, Y. Huang, Z. Lin, M. Lei, Y. Xue, L. Guo, and C. Zhou, "Asymmetric motion vector refinement for future video coding," in *2024 Data Compression Conference (DCC)*. IEEE, 2024, pp. 402–411.

[20] H. Huang, Z. Zhang, V. Seregin, W.-J. Chien, C.-C. Chen, and M. Karczewicz, "Adaptive bilateral matching for decoder-side motion vector refinement in video coding," in *2022 Data Compression Conference (DCC)*. IEEE, 2022, pp. 01–07.

[21] J. Chen, R.-L. Liao, Y. Ye, and X. Li, "Decoder-side affine model refinement for video coding beyond vvc," in *2023 Data Compression Conference (DCC)*. IEEE, 2023, pp. 248–257.

[22] L. Li, H. Li, D. Liu, Z. Li, H. Yang, S. Lin, H. Chen, and F. Wu, "An efficient four-parameter affine motion model for video coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 8, pp. 1934–1948, 2017.