Deep Affine Motion Compensation Network for Inter Prediction in VVC

Dengchao Jin, Jianjun Lei, *Senior Member, IEEE*, Bo Peng, *Member, IEEE*, Wanqing Li, *Senior Member, IEEE*, Nam Ling, *Life Fellow, IEEE*, and Qingming Huang, *Fellow, IEEE*

***Abstract*— In video coding, it is a challenge to deal with scenes with complex motions, such as rotation and zooming. Although affine motion compensation (AMC) is employed in Versatile Video Coding (VVC), it is still difficult to handle non-translational motions due to the adopted hand-craft block-based motion compensation. In this paper, we propose a deep affine motion compensation network (DAMC-Net) for inter prediction in video coding to effectively improve the prediction accuracy. To the best of our knowledge, our work is the first attempt to deal with the deformable motion compensation based on CNN in VVC. Specifically, a deformable motion-compensated prediction (DMCP) module is proposed to compensate the current encoding block through a learnable way to estimate accurate motion fields. Meanwhile, the spatial neighboring information and the temporal reference block as well as the initial motion field are fully exploited. By effectively fusing the multi-channel feature maps from DMCP, an attention-based fusion and reconstruction (AFR) module is designed to reconstruct the output block. The proposed DAMC-Net is integrated into VVC and the experimental results demonstrate that the proposed method considerably enhances the coding performance.**

***Index Terms*— Video coding, VVC, affine motion compensation, deep neural network, deformable motion compensation.**

I. Introduction

With the prevalence of high-definition (HD) and ultra-high-definition (UHD) videos, the demand for high efficiency video compression techniques has increased dramatically. The ITU-T Video Coding Experts Group (VCEG) and ISO/IEC Moving Picture Experts Group (MPEG) have developed a series of video compression coding standards [1], [2]. Different from image coding, video coding generally focuses on removing temporal redundancy by inter prediction with motion compensation to effectively boost coding performance. In the process of motion compensation, pixels of each block are first predicted with the highest similar block in reference frames, and then the residual between predicted pixels and real pixels is encoded into bitstream. Therefore, how to improve the prediction accuracy of motion compensation is highly critical for boosting compression efficiency.

In the latest Versatile Video Coding (VVC), the existing translational motion compensation (TMC) and advanced affine motion compensation (AMC) [3] are jointly exploited for eliminating temporal redundancy. TMC predicts pixels on the assumption that movement between video frames is translational. Therefore, non-translational motions in natural videos will result in a large residual in TMC. To address this issue, AMC is integrated into VVC to improve the ability to deal with complex motions. Although AMC has significantly improved coding performance, there still exist several limitations. First, the subblock-wised motion field is derived from fixed points by hand-craft algorithms, thus resulting in blocking artifacts between sub-blocks, and inaccurate prediction in some high-order motions, such as bilinear and perspective motions. Second, existing AMC algorithms pay more attention to the correlation in temporal domain, while the spatial neighboring information in the current frame is not effectively utilized.

In the past years, deep learning-based methods have achieved promising results in several image and video processing tasks, such as classification, super-resolution, and attention prediction [4]–[8]. Inspired by the success of deep learning, recent researches have been devoted to developing learning-based tools for traditional video coding schemes [9]–[25] and learning-based end-to-end compression schemes [26]–[31]. Specifically, several studies [9]–[12] have attempted to substitute or enhance TMC with convolutional neural networks (CNNs) to improve the coding performance. However, there is no report yet on CNN-based AMC to effectively deal with complex motions.

This paper proposes a deep affine motion compensation network (DAMC-Net) to boost the performance of AMC. The main idea of the proposed method is to compensate the current encoding block by estimating accurate motion fields

Manuscript received April 30, 2021; revised July 27, 2021; accepted August 10, 2021. Date of publication August 24, 2021; date of current version June 6, 2022. The work of Jianjun Lei and Bo Peng was supported in part by the National Key R&D Program of China under Grant 2018YFE0203900, in part by the National Natural Science Foundation of China under Grant 61931014 and Grant 61722112, and in part by the Natural Science Foundation of Tianjin under Grant 18JCJQJC45800. The work of Qingming Huang was supported in part by the National Natural Science Foundation of China under Grant 61620106009 and Grant U1636214. This article was recommended by Associate Editor G. Correa. *(Corresponding author: Jianjun Lei.)*

Dengchao Jin, Jianjun Lei, and Bo Peng are with the School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China (e-mail: jdc3159761141@tju.edu.cn; jjlei@tju.edu.cn; bpeng@tju.edu.cn).

Wanqing Li is with the Advanced Multimedia Research Laboratory, University of Wollongong, Wollongong, NSW 2522, Australia (e-mail: wanqing@uow.edu.au).

Nam Ling is with the Department of Computer Science and Engineering, Santa Clara University, Santa Clara, CA 95053 USA (e-mail: nling@scu.edu).

Qingming Huang is with the School of Computer and Control Engineering, University of Chinese Academy of Sciences, Beijing 100190, China (e-mail: qmhuang@ucas.ac.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSVT.2021.3107135.

Digital Object Identifier 10.1109/TCSVT.2021.3107135

with learning-based methods. In addition, the multi-domain information from the current reconstructed frame and the temporal reference frame as well as the initial motion field are fully exploited to provide informative patches. Different from existing AMC in VVC, motion fields in DAMC-Net are learnable. Thus, the proposed method is capable of dealing with most complex motions in natural videos. The major contributions of this paper are summarized as follows.

- Aiming to boost the performance of inter prediction in video coding, a DAMC-Net is proposed to improve the prediction accuracy of AMC. To the best of our knowledge, the proposed DAMC-Net is the first attempt to perform deformable prediction task based on CNNs in VVC.
- A deformable motion-compensated prediction (DMCP) module is designed to compensate the current encoding block by estimating accurate motion fields with multi-domain information as references.
- The proposed DAMC-Net is integrated into VVC and experimental results in both affine inter-mode and affine merge-mode have demonstrated that the proposed method significantly increases the selection rate of affine modes and subsequently achieves considerable coding performance improvement.

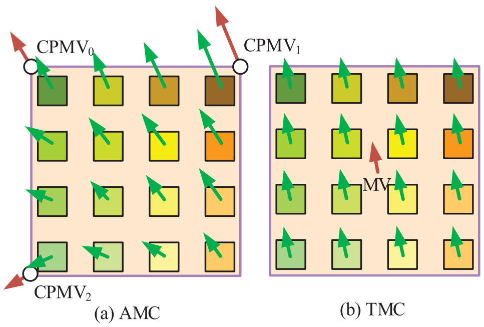
The rest of this paper is organized as follows. Section II reviews the related work. Section III introduces the proposed method in detail. Experimental results are shown in Section IV. Finally, Section V concludes the paper.

II. Related Work

A. Affine Motion Compensation in VVC

For the current encoding block, motion compensation generally obtains the most similar block among the reference frames as its prediction signal. By integrating TMC into video coding standards, such as High Efficiency Video Coding (HEVC), the relatively accurate motion fields of most blocks are estimated. However, it is difficult to further improve the coding performance of TMC due to its limited ability to model complex motions. With an increasing demand for video compression efficiency, more sophisticated models are needed to handle complex motions. To this end, Joint Video Exploration Team (JVET) integrated both 4-parameter [3] and 6-parameter affine motion models of AMC into VVC as an inter-prediction tool [32], [33].

The difference between AMC and TMC is shown in Fig. 1. For each 4 × 4 sub-block in the current block, AMC utilizes multiple Control Points Motion Vectors (CPMVs) to derive specific Motion Vector (MV), while TMC utilizes a common MV of the current block to represent the MVs for all sub-blocks. Therefore, AMC has the ability to characterize complex motions more effectively than TMC. In VVC, there are two kinds of affine motion models for AMC, *i.e.*, 4-parameter affine model and 6-parameter affine model. Specifically, fewer bits are required to signal the CPMVs of a 4-parameter affine model, and more complex motions can be represented by a 6-parameter affine model. Besides, affine inter-mode and affine merge-mode are applied as two inter prediction modes in VVC. For the affine inter-mode, CPMVs are determined in the encoder and signalled explicitly to the decoder. For the affine merge-mode, CPMVs are derived implicitly from neighboring blocks.



(a) AMC (b) TMC

Fig. 1. Difference between (a) AMC and (b) TMC.

The above traditional affine motion compensation generally predicts the current block by a parameterized affine model, which is hard to effectively deal with irregular motions. In this paper, the proposed DAMC-Net is a pixel-wise motion model, which is much more flexible than a 4-parameter affine model and a 6-parameter affine model, hence, its robustness for scenes with complex motions.

B. Deep Learning for Inter Prediction

Inspired by the success of deep learning, many researchers have employed CNNs to improve the performance of video coding, and achieved superior coding efficiency in filtering, intra prediction, and inter prediction.

For the learning-based tools for inter prediction of traditional coding schemes, several works have been proposed to refine or substitute modules of uni-directional and bi-directional prediction in HEVC. In [9] and [12], CNN-based methods were proposed to refine the uni-directional prediction of TMC in HEVC. To improve coding performance of bi-directional prediction, Zhao *et al.* [10] proposed a fully-convolutional neural network to learn a mapping between bi-directional compensated blocks and final prediction signals. Mao and Yu [11] took spatial pixels and temporal display orders as additional information to improve the accuracy of bi-directional prediction. Yan *et al.* [13] proposed a fractional-pixel reference generation network FRCNN to perform fractional-pixel motion compensation. Inspired by the works on video prediction, there were also several works focusing on directly extrapolating or interpolating the prediction of the current frame as an additional reference. For instance, Lin *et al.* [14] proposed a video coding oriented Laplacian pyramid of generative adversarial networks (VC-LAPGAN) to predict the current frame. Huo *et al.* [15] combined block-based motion-compensated prediction and frame extrapolation to generate an additional reference frame, which achieves considerable performance improvement in both HEVC and VVC. Zhao *et al.* [16] proposed a novel network to generate a high-quality reference and devised a CTU level coding tool to achieve a trade-off between performance and complexity. Xia *et al.* [17] proposed a multi-scale network to

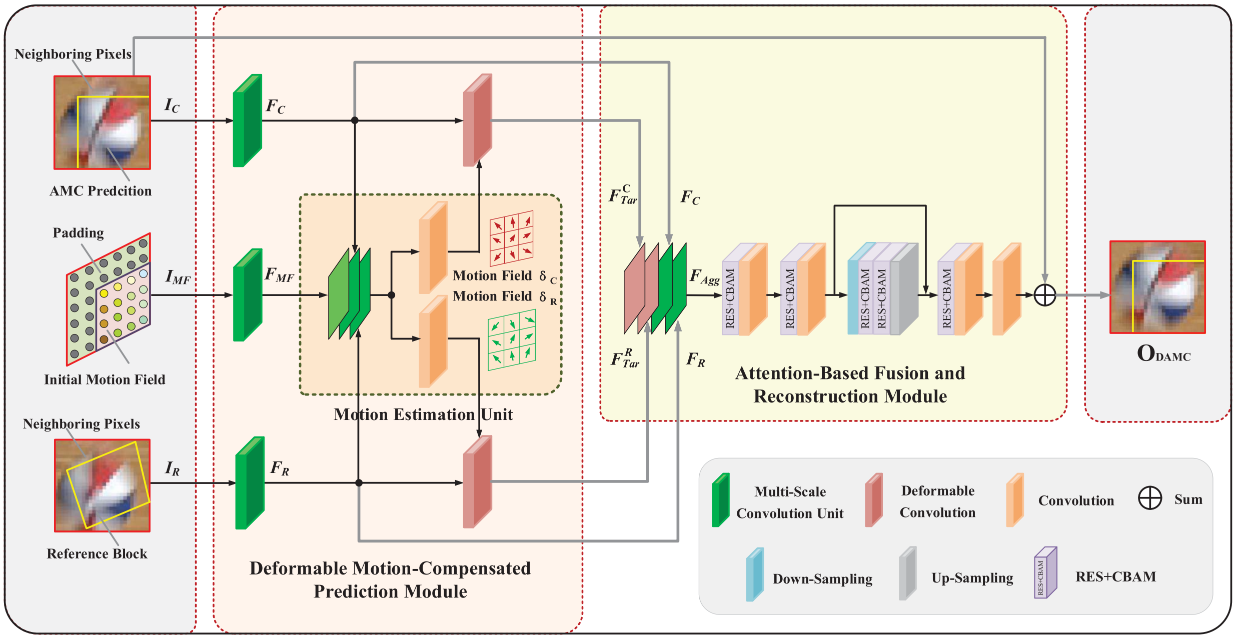


Fig. 2. Overall architecture of the proposed method.

generate an additional reference frame from coarse to fine. Liu *et al.* [18] proposed a multi-scale quality attentive factorized kernel convolutional neural network (MQ-FKCNN) to synthesize an additional reference frame. Choi and Bajic [19] utilized both decoded frames and temporal indices to generate a reference frame for video coding scheme. They also proposed an affine transformation-based scheme [20], in which the spatially-varying filters and affine parameters are computed to generate the warped samples for synthesizing the reference frame.

However, to the best of our knowledge, no work has been reported on learning-based AMC. Moreover, the existing learning-based tools for TMC pay little attention to estimating the complex motion field for compensating the current block, while estimating the accurate motion field is essential for motion-compensated prediction.

III. The Proposed Method

In this section, the proposed method is presented in detail. First, the architecture of the proposed method is systematically introduced. Second, the deformable motion-compensated prediction module which plays an important role in the proposed network is illustrated. Third, the attention-based fusion and reconstruction module is illustrated. Finally, the details of integrating DAMC-net into VVC are described.

A. Architecture of DAMC-Net

Traditional AMC compensates the current block merely by a parameterized affine model, which has limited capability to model complex motions. To solve this problem, a learning-based model, DAMC-Net, is designed to compensate the current block by explicitly estimating pixel-wise motion fields rather than implicitly deriving the subblock-wise motion field. Specifically, a spatial pixel-wise motion field is estimated to refine the prediction block for alleviating the spatial blocking artifacts, and a temporal pixel-wise motion field between frames is estimated to compensate the current block for alleviating temporal misalignment.

Fig. 2 shows the overall architecture of the proposed DAMC-Net. As shown in the figure, the multi-domain information is fully leveraged in the proposed DAMC-Net. In order to improve the prediction accuracy of AMC, the spatial neighboring pixels of the current block as well as its prediction of AMC are combined as the first input (I_C) to explore spatial correlations. Besides, to obtain as accurate as possible source pixels in the temporal reference frame of the current block, the most similar block together with neighboring pixels in the reference frame is constructed based on CPMVs and used as the second input (I_R). More importantly, since the initial motion field (I_{MF}) constructed by CPMVs contains motion information, I_{MF} is utilized as the third input. Taking the I_C, I_{MF}, and I_R as the inputs, the proposed DAMC-Net is optimized with respect to jointly utilizing spatial neighboring and temporal correlative information to improve the prediction accuracy.

In the DMCP module, features F_C, F_{MF}, and F_R are first extracted from I_C, I_{MF}, and I_R respectively. Taking these features as inputs, the motion estimation unit (MEU) is designed to estimate motion fields. Based on estimated motion fields, deformable convolution is used to compensate F_C and F_R. Features of compensated output F^C_{Tar} and F^R_{Tar} are concatenated with F_C as well as F_R to construct the aggregated feature F_{Agg}. Finally, the output block, O_{DAMC}, is reconstructed from F_{Agg} by an AFR module.

B. Deformable Motion-Compensated Prediction (DMCP)

Due to the limitation of deriving the subblock-wised motion field, the prediction of AMC suffers from misalignment at pixel-level. In order to improve the granularity of AMC, the DMCP module is designed to estimate pixel-wise motion fields for compensating the current block.

In DMCP, to extract deep features with abundant information, features F_C, F_{MF}, and F_R are extracted from I_C, I_{MF}, and I_R by multi-scale convolution unit [25] respectively. Then, the MEU is designed to estimate accurate motion fields. As shown in Fig. 2, F_C, F_{MF}, and F_R are first concatenated, then separate convolution operations are followed to generate offsets for each texture branch in MEU. It should be noticed that not only the texture information I_C and I_R are exploited in MEU, but also the initial motion field I_{MF} are jointly utilized to estimate accurate motion fields. Compared to a network which learns motion fields from scratch, the DMCP-Net estimates accurate motion fields with coarse input, which helps to reduce the network training difficulty and ensure the quality of learned motion fields.

Let δ_C denote the motion field from F_C to F_{Tar}^C, which computes the affine motion between I_C and O_{DAMC}. δ_R denotes the motion field from F_R to F_{Tar}^R, which computes the affine motion between I_R and O_{DAMC}. Since the motion between I_C and O_{DAMC} is smaller than that between I_R and O_{DAMC}, the MEU actually estimates a fine motion field from I_C to O_{DAMC}. Calculation of motion fields δ_C and δ_R for the two texture branches can be expressed as:

$$\begin{cases} \delta_C = \mathcal{F}_{\theta 1}\left(F_C, F_{MF}, F_R\right) \\ \delta_R = \mathcal{F}_{\theta 2}\left(F_C, F_{MF}, F_R\right) \end{cases} \tag{1}$$

where $\theta 1$ and $\theta 2$ are parameters learned by the network. $\mathcal{F}(\cdot)$ represents the operation of the motion estimation unit.

Similar to the function of the affine motion compensation module in VVC, features F_C and F_R of the two texture branches are deformed separately to generate compensated features. Inspired by [34], motion compensation is operated by deformable convolution, which adaptively deforms the kernel sampling under the control of a motion field. Therefore, compensated features F_{Tar}^C and F_{Tar}^R for the two texture branches can be computed as follows.

$$\begin{cases} F_{Tar}^R = DConv\left(F_R, \delta_R\right) \\ F_{Tar}^C = DConv\left(F_C, \delta_C\right) \end{cases} \tag{2}$$

where $DConv(\cdot)$ denotes the deformable convolution [34]. Since DMCP compensates the feature maps rather than the pixels of the target image, it effectively exploits non-local context.

C. Attention-Based Fusion and Reconstruction (AFR)

Taking the outputs of the DMCP module as input, the AFR module fuses the multi-channel information and reconstructs the final prediction signal. In order to obtain feature representation with abundant information and improve the quality of the final prediction signal, the AFR module obtains the aggregated feature F_{Agg} by fusing the non-deformed features F_C and F_R with the compensated features F_{Tar}^C and F_{Tar}^R. Due to multiple sources of information are included in F_{Agg}, an attention mechanism is employed to emphasize useful features and suppress the others. Considering that the residual block can extract deep features effectively, the Res + CBAM structure [35], being composed of CBAM and residual block, is adopted. Furthermore, a down-sampling layer and an up-sampling layer are employed to increase the receptive field and preserve the low frequency information, with two Res + CBAM units between them. The skip connection [36] is designed to accelerate training process. As shown in Fig. 2, the overall framework of the AFR module stacks five Res + CBAM units. To optimize the proposed network, $\mathcal{L}_2$ loss is utilized as the loss function:

$$\mathcal{L} = \|\left(O_{GT} - O_{DAMC}\right)\|_2^2 \tag{3}$$

where O_{GT} is the corresponding block in the raw videos and O_{DAMC} is the output of AFR.

D. Integration of DAMC-Net in VVC

1) The Scope of DAMC-Net in VVC: There are two affine modes in VVC, namely, affine inter-mode and affine merge-mode. Since these two modes are both based on AMC, the proposed DAMC-Net is applied to these two modes. In addition, since the flexible quadtree nested multi-type tree structure is explored in VVC to split coding tree unit (CTU) into CUs, there exist CUs with square or rectangular shape in VVC. Overall, there are 12 sizes of CUs with affine inter-mode, and 19 sizes of CUs with affine merge-mode. In order to ensure the performance of DAMC-Net in each size of CU, a series of models corresponding to different sizes of CUs are exploited in this paper. Therefore, 12 models are trained for the affine inter-mode and 19 models for the affine merge-mode. The following section illustrates the performance of two affine modes integrated with DAMC-Net.

2) The Strategy of DAMC-Net in VVC: The DAMC-Net is defined as a new DAMC mode and embedded into the process of CU optimal mode decision. The DAMC mode is utilized as an optional mode for the CUs with affine inter-mode and affine merge-mode and determined whether to be selected in the encoder based on rate-distortion optimization (RDO). As for the DAMC mode in the affine inter-mode, DAMC-Net is first fed with inputs after AMC, and outputs the compensated block O_{DAMC}. Then, the RDO process determines whether to select the DAMC mode, and a designed flag recording the decision result of the RDO is signalled to decoder. As for the DAMC mode in the affine merge-mode, considering the encoding complexity, it is determined whether to be selected by the RDO after the best affine merge candidate is searched. Meanwhile, only the CUs with the affine merge-mode need to encode and decode the flag. Since the DAMC-Net is nested in the process of CU optimal mode decision rather than post-processing after encoding frames, the selection rate of CUs with affine modes increases significantly.

IV. Experiments

A. Experimental Settings

1) Training Data Preparation: To evaluate the proposed DAMC-Net, a training dataset is first collected with 106 videos at different resolutions from [37], [38] and 8 videos at the resolution of 3840 × 2160 from [39], that is 114 raw video sequences with rich and complex motions in total. Considering the complexity of VVC, the 4K videos are down-sampled to 1280 × 720. Then, VTM-6.2 is used to compress the

TABLE I
BD-RATE RESULTS OF THE PROPOSED DAMC-NET FOR AFFINE INTER-MODE

Class	Sequence	Inter AMC			Inter DAMC-Net		
		Y	Cb	Cr	Y	Cb	Cr
Class B	MarketPlace	-5.27%	-3.77%	-7.24%	-5.36%	-3.76%	-6.74%
	RitualDance	-1.69%	-1.13%	-1.20%	-1.89%	-1.34%	-1.31%
	Cactus	-8.43%	-6.06%	-6.24%	-8.88%	-6.30%	-6.27%
	BasketballDrive	-2.45%	-2.11%	-2.18%	-2.68%	-2.36%	-2.01%
	BQTerrace	-0.77%	-0.24%	-0.64%	-1.74%	-0.72%	-0.96%
	Average Class B	-3.72%	-2.66%	-3.50 %	-4.11%	-2.90%	-3.46%
Class C	RaceHorses	-1.39%	-1.04%	-1.02%	-1.63%	-1.33%	-0.23%
	BQMall	-0.88%	-0.38%	0.17%	-1.20%	-0.48%	-0.03%
	PartyScene	-1.52%	-1.21%	-1.12%	-2.37%	-1.83%	-1.73%
	BasketballDrill	-1.04%	-0.20%	-0.46%	-1.16%	-0.23%	-0.07%
	Average Class C	-1.21%	-0.71%	-0.61 %	-1.59%	-0.97%	-0.52%
Class D	RaceHorses	-1.79%	-0.83%	-0.40%	-2.14%	-1.03%	-0.79%
	BQSquare	-3.43%	-1.21%	-2.66%	-6.13%	-2.95%	-4.27%
	BlowingBubbles	-1.88%	-0.68%	-1.66%	-2.62%	-1.18%	-1.62%
	BasketballPass	-1.74%	-1.77%	-0.32%	-2.81%	-2.47%	0.86%
	Average Class D	-2.21%	-1.12%	-1.10 %	-3.43%	-1.91%	-1.46%
Class E	FourPeople	-0.75%	-0.22%	-0.57%	-1.27%	-0.65%	-0.66%
	Johnny	-2.76%	-1.58%	-1.72%	-4.39%	-2.61%	-2.36%
	KristenAndSara	-3.18%	-2.27%	-2.50%	-3.41%	-2.31%	-2.99%
	Average Class E	-2.23%	-1.36%	-1.60 %	-3.02%	-1.86%	-2.00%
Overall		**-2.44%**	-1.54%	-1.82%	**-3.11%**	-1.97%	-1.95%

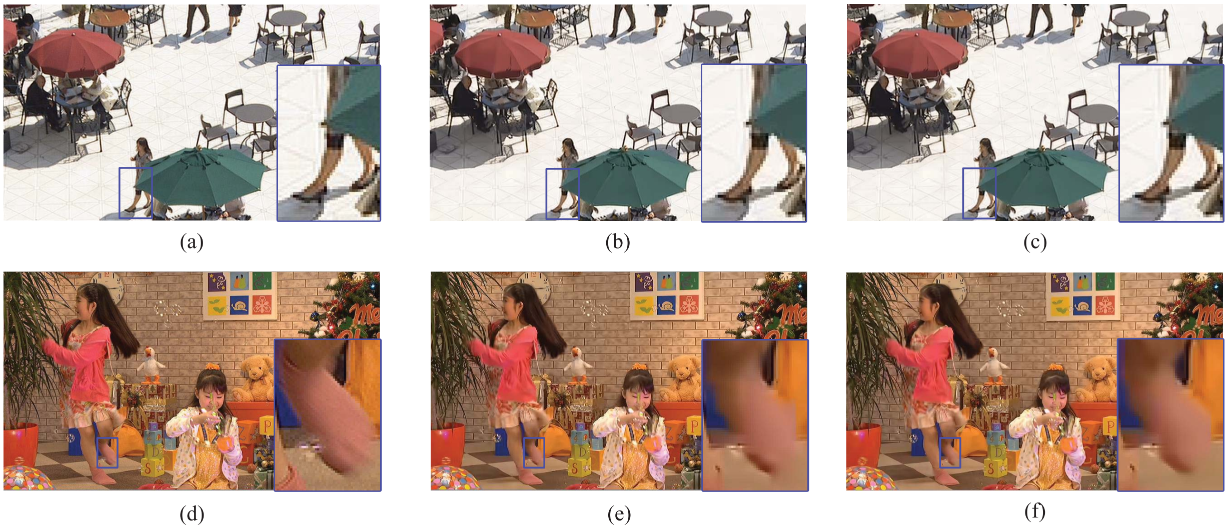

(a)
(b)
(c)
(d)
(e)
(f)

Fig. 3. Visual comparison between VTM-6.2 and the proposed Inter & Merge DAMC-Net. Top: The 3-rd frame of BQsquare under QP 27. Bottom: The 30-th frame of PartyScene under QP 32. (a) BQsquare original image, (b) Inter & Merge AMC (5216 bits, 35.24 dB), (c) Proposed (**4440 bits**, 35.42 dB), (d) PartyScene original image, (e) Inter & Merge AMC (22408 bits, 29.82 dB), (f) Proposed (**21984 bits**, 29.91 dB).

video sequences configured with Low Delay P (LDP) under four quantization parameters (QPs) $\{22, 27, 32, 37\}$. Due to the high similarity between adjacent frames, video frames are sampled at the regular interval of 3 to generate the training samples. In the process of compression, I_C, I_{MF}, and I_R of CUs with affine modes in the selected frames, and the

TABLE II
BD-RATE RESULTS OF THE PROPOSED DAMC-NET FOR BOTH AFFINE INTER-MODE AND AFFINE MERGE-MODE

Class	Sequence	Inter & Merge AMC			Inter & Merge DAMC-Net		
		Y	Cb	Cr	Y	Cb	Cr
Class B	MarketPlace	-5.28%	-3.81%	-4.94%	-5.60%	-3.51%	-4.13%
	RitualDance	-1.82%	-1.29%	-1.56%	-2.45%	-2.49%	-1.58%
	Cactus	-9.16%	-7.07%	-7.21%	-10.83%	-7.73%	-8.02%
	BasketballDrive	-2.80%	-2.45%	-2.93%	-3.52%	-3.04%	-2.76%
	BQTerrace	-0.85%	-0.45%	-0.73%	-3.57%	-1.42%	-0.98%
	Average Class B	-3.98%	-3.01%	-3.47 %	-5.19%	-3.64%	-3.49%
Class C	RaceHorses	-1.44%	-0.65%	-1.38%	-1.87%	-0.63%	-0.55%
	BQMall	-1.42%	-1.11%	-0.52%	-2.70%	-1.04%	-0.16%
	PartyScene	-1.83%	-1.37%	-1.48%	-3.77%	-2.00%	-1.63%
	BasketballDrill	-1.49%	-1.22%	-1.48%	-2.44%	-1.42%	-1.38%
	Average Class C	-1.55%	-1.09%	-1.22 %	-2.70%	-1.27%	-0.93%
Class D	RaceHorses	-1.44%	-1.45%	-0.58%	-2.50%	-4.02%	-1.38%
	BQSquare	-4.10%	-2.34%	-3.19%	-8.91%	-5.67%	-5.99%
	BlowingBubbles	-2.15%	-0.58%	-0.91%	-3.78%	-1.42%	-1.54%
	BasketballPass	-2.90%	-3.20%	-2.37%	-4.41%	-2.64%	3.06%
	Average Class D	-2.65%	-1.89%	-1.76 %	-4.90%	-3.44%	-2.99%
Class E	FourPeople	-1.79%	-1.36%	-1.86%	-2.98%	-1.76%	-1.99%
	Johnny	-3.27%	-1.47%	-1.94%	-4.76%	-2.42%	-3.10%
	KristenAndSara	-4.59%	-3.94%	-3.63%	-5.07%	-3.38%	-3.78%
	Average Class E	-3.22%	-2.26%	-2.48 %	-3.93%	-2.52%	-2.96%
Overall		**-2.90%**	-2.11%	-2.29%	**-4.32%**	-2.79%	-2.63%

corresponding ground-truth in raw video frames are utilized to construct training samples. Consequently, 76 sub-datasets in total are obtained, which correspond to 4 QPs and 19 CU sizes (8 × 8, 8 × 16, 8 × 32, 8 × 64, 16 × 8, 16 × 16, 16 × 32, 16 × 64, 32 × 8, 32 × 16, 32 × 32, 32 × 64, 64 × 8, 64 × 16, 64 × 32, 64 × 64, 64 × 128, 128 × 64, 128 × 128).

2) Encoding Configurations: DAMC-Net is integrated into VVC reference software VTM (version 6.2). Experiments are performed under the JVET common test conditions (CTC) [40]. Since single reference frame is utilized in the proposed DAMC-Net, LDP configuration and Classes B∼E are tested. In the experiments, the testing QPs are set as $\{22, 27, 32, 37\}$, and the widely employed BD-rate [41], [42] is used as the objective metric to evaluate the coding performance. A CPU + GPU cluster is used as the test environment, where VVC coding is tested in CPU and the DAMC-Net is running in GPU. The CPU is Inter(R) Core(TM) i9-9900K CPU @ 3.60GHz, and the GPU is NVIDIA GeForce GTX 1080Ti.

3) Training Strategy: The proposed DAMC-Net is implemented on TensorFlow [43] and trained on a Nvidia GeForce GTX 1080Ti GPU. To satisfy all the sizes of CUs with affine mode, 12 models for affine inter-mode and 19 models for affine merge-mode are trained for each QP. In the training phase, the base model with the affine inter-mode, QP of 22, and CU size of 16 × 16, is firstly trained. Specifically, the network is optimized using Adam [44] with a batch size of 128, and the learning rate is initially set to 0.0001 for the first 2,000,000 steps and decayed to 0.00001 for the last 1,000,000 steps. Then, other models are refined based on the base model with the learning rate of 0.00005 for 100,000 steps.

B. Comparison Results and Analyses

To validate the effectiveness of the proposed DAMC-Net, performance of the scheme of DAMC-Net for affine inter-mode (Inter DAMC-Net) is first compared with the scheme of VTM-6.2 with affine inter-mode (Inter AMC). Meanwhile, the VTM-6.2 without affine inter-mode is set as the baseline to compute BD-rate. The coding performance on Y, Cb, and Cr components of different methods are reported in Table I. As shown in the table, on the Y component, the "Inter DAMC-Net" achieves 4.11%, 1.59%, 3.43%, and 3.02% BD-rate reduction on average for Class B, C, D, and E, respectively. Particularly, the "Inter DAMC-Net" achieves up to 6.13% BD-rate reduction on *BQsquare*, while "Inter AMC" only obtains 3.43% BD-rate reduction. Besides, to further validate the advantage of the DAMC-Net, experiments about

TABLE III
SELECTION RATES OF THE PROPOSED DAMC MODE

Class	Sequences	**Inter DAMC-Net**	
		HR	IR
Class B	MarketPlace	72.16%	208.50%
	RitualDance	68.04%	154.01%
	Cactus	80.37%	422.91%
	BasketballDrive	82.06%	274.43%
	BQTerrace	82.20%	1297.30%
	Average Class B	76.97%	471.43%
Class C	RaceHorces	74.13%	310.36%
	BQMall	71.35%	232.94%
	PartyScene	85.87%	314.69%
	BasketballDrill	71.21%	231.11%
	Average Class C	75.64%	272.28%
Class D	RaceHorses	77.88%	190.53%
	BQSquare	61.47%	436.56%
	BlowingBubbles	83.31%	380.63%
	BasketballPass	73.01%	293.94%
	Average Class D	73.92%	325.41%
Class E	FourPeople	70..44%	235.81%
	Johnny	59.78%	264.59%
	KristenAndSara	74.02%	218.32%
	Average Class E	68.08%	239.57%
Average		**74.21%**	**341.66%**

TABLE IV
SELECTION RATES OF THE DAMC MODE BASED ON THE AREA

Class	Sequences	**Inter DAMC-Net**		**Inter & Merge DAMC-Net**	
		Ours	Inter AMC + Ours	Ours	AMC + Ours
Class B	MarketPlace	22.47%	45.48%	37.15%	53.86%
	RitualDance	4.37%	10.26%	11.92%	29.98%
	Cactus	9.23%	11.18%	33.01%	56.31%
	BasketballDrive	5.19%	7.92%	18.78%	34.59%
	BQTerrace	4.16%	6.31%	26.78%	41.66%
	Average Class B	9.08%	16.23%	25.53%	43.28%
Class C	RaceHorces	7.33%	10.36%	14.77%	26.06%
	BQMall	3.08%	4.43%	12.92%	28.68%
	PartyScene	4.91%	6.56%	20.44%	33.61%
	BasketballDrill	3.67%	5.48%	8.27%	14.63%
	Average Class C	4.75%	6.71%	14.10%	25.75%
Class D	RaceHorses	5.89%	10.46%	14.46%	27.52%
	BQSquare	6.07%	8.25%	24.89%	42.72%
	BlowingBubbles	10.25%	14.62%	21.68%	39.62%
	BasketballPass	2.02%	3.07%	6.72%	30.65%
	Average Class D	6.06%	9.10%	16.94%	35.13%
Class E	FourPeople	2.85%	3.75%	10.58%	33.04%
	Johnny	6.81%	7.94%	11.38%	19.88%
	KristenAndSara	6.79%	8.23%	13.74%	32.57%
	Average Class E	5.48%	6.64%	11.90%	28.50%
Overall		**6.57%**	**10.27%**	**17.97%**	**34.09%**

the scheme of DAMC-Net for both affine inter-mode and affine merge-mode (Inter & Merge DAMC-Net) are conducted. Since the proposed DAMC mode is set as an optional method for all the CUs with affine mode, the VTM-6.2 without both affine inter-mode and affine merge-mode is set as the baseline for a fair comparison. The comparison results are shown in Table II, it can be seen that the proposed “Inter & Merge DAMC-Net” achieves significant bits saving than the “Inter & Merge AMC”. Specifically, 4.32% BD-rate reduction on average is achieved in the proposed scheme of “Inter & Merge DAMC-Net” on Y component. Compared with the scheme of “Inter & Merge AMC”, the proposed method further achieves 50% coding gain in average and even doubles the coding gain on several sequences, such as *BQTerrace* in Class B and *PartyScene* in Class C, which strongly demonstrates the effectiveness of the proposed DAMC-Net. Especially, on sequences with complex motions, such as *Cactus*, the proposed method achieves more coding gain than the average, which demonstrates that the proposed DAMC-Net is effective in dealing with the complex motions in these scenes.

In addition, in order to intuitively verify the effectiveness of the proposed method, the visual comparison of the decoded images obtained by different methods are shown in Fig. 3. It can be observed that the proposed DAMC-Net obtains the decoded images with higher quality while achieving more BD-rate reduction. Moreover, the pixels on the boundaries of the moving objects obtained by the proposed method are more closing to the raw videos, due to the explicit motion fields estimated for motion compensation.

C. Mode Selection Results

To further analyze the contribution of the proposed method, the results of mode selection of the DAMC mode for affine inter mode on all the test sequences with QP 22 are reported in Table III. Let N^O_{AMC} denote the number of CUs with the affine inter-mode in the original VTM, N^P_{AMC} denote the number of CUs with the affine inter-mode in the proposed scheme with “Inter DAMC-Net”, and N_{DAMC} denote the number of CUs with the DAMC mode in “Inter DAMC-Net”. As shown in Table III, the selection rate HR of DAMC mode in “Inter DAMC-Net” and increasing rate IR of affine inter-mode are defined as follows.

$$HR = N_{DAMC}/(N_{DAMC} + N^P_{AMC}) \tag{4}$$

$$IR = (N_{DAMC} + N^P_{AMC})/N^O_{AMC} \tag{5}$$

It can be observed that the proposed DAMC mode is selected for most cases with the HR of 74%. More importantly, larger IR is obtained by utilizing the proposed

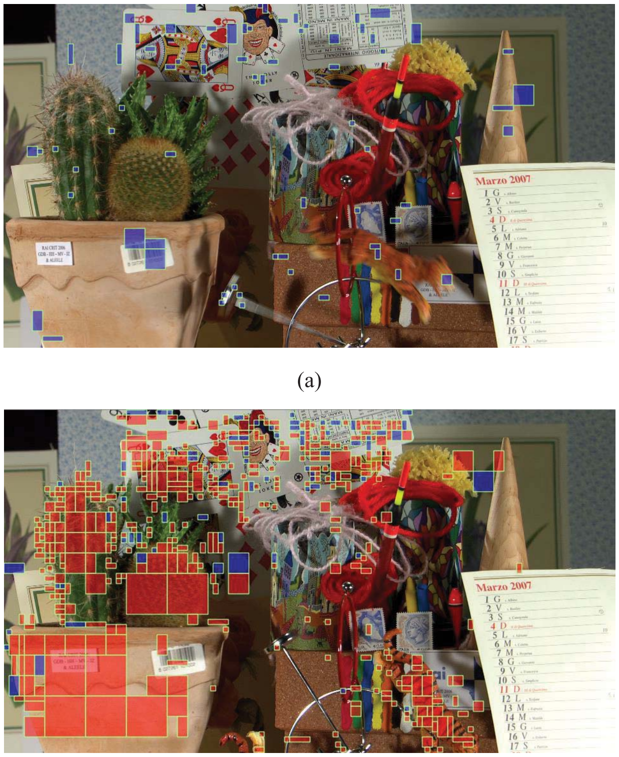
TABLE V
BD-Rate Results of the DAMC-Net Without DMCP Module

Class	Sequence	Without DMCP		
		Y	Cb	Cr
Class B	MarketPlace	-5.16%	-3.72%	-7.25%
	RitualDance	-1.67%	-1.72%	-1.38%
	Cactus	-8.80%	-6.44%	-6.84%
	BasketballDrive	-2.59%	-2.33%	-1.82%
	BQTerrace	-0.80%	-0.13%	-0.20%
	Average Class B	-3.81%	-1.89%	-3.50%
Class C	RaceHorses	-1.70%	-1.58%	-0.83%
	BQMall	-0.91%	-0.48%	0.03%
	PartyScene	-1.61%	-1.03%	-1.24%
	BasketballDrill	-1.17%	-0.29%	0.05%
	Average Class C	-1.35%	-0.85%	-0.50%
Class D	RaceHorses	-2.13%	-1.27%	-0.76%
	BQSquare	-4.31%	-2.64%	-3.13%
	BlowingBubbles	-2.19%	-1.50%	-2.11%
	BasketballPass	-2.62%	-1.53%	-0.43%
	Average Class D	-2.81%	-1.74%	-1.61%
Class E	FourPeople	-1.06%	-0.26%	-0.47%
	Johnny	-3.37%	-2.74%	-2.68%
	KristenAndSara	-3.49%	-2.51%	-3.29%
	Average Class E	-2.64%	-1.84%	-2.15%
Overall		**-2.72%**	-1.89%	-2.02%

DAMC-Net. It is mainly because that the affine inter-mode is more likely to be selected, as the proposed DAMC-Net effectively eliminates the negative effects of the hand-craft algorithms and improves the performance of AMC. Besides, higher HR and IR are obtained on the sequences with complex motions, such as *Cactus* and *PartyScene*, which further demonstrates the effectiveness of the DAMC-Net than traditional AMC in dealing with complex motions. In addition, it can also be seen that high-definition sequences in Class B have the highest HR and IR among all test sequences, which indicates that the proposed method is more suitable for high-definition videos.

Meanwhile, the results of mode selection based on the influenced area on all test sequences with QP 22 are also reported in Table IV. It can be observed that the proposed DAMC mode is more likely to be selected than the traditional AMC for most cases. Especially, higher selection rates on the sequences with complex motions, *MarketPlace* and *PartyScene*, also demonstrate the effectiveness of the proposed DAMC-Net in dealing with complex motions.

Fig. 4 (a) shows the mode selection result of VTM-6.2 with affine inter-mode. The highlighted blocks are the CUs selected with affine inter-mode. Fig. 4 (b) shows the mode selection result of the proposed scheme with "Inter DAMC-Net". The red block denotes the CUs with DAMC mode, while the blue block denotes the CUs with affine inter-mode. It is obvious that the DAMC mode is more likely to be selected in the CUs with complex motions, and the selection rate of affine modes increases significantly.


(a)

(b)

Fig. 4. Mode selection results for the 21-st frame of *Cactus* under QP 22. (a) Inter AMC. (b) Proposed.

D. Ablation Study

The DMCP module is designed to compensate the current encoding block by estimating more accurate motion fields. In order to illustrate the contribution of the DMCP module, DMCP module is removed from DAMC-Net, and only F_C is taken as the input of AFR. The results of the proposed method without the DMCP module on the scheme with "Inter DAMC-Net" are shown in Table V. Compared with Tabel I, the BD-rate reduction on average decreases from 4.11%, 1.59%, 3.43%, and 3.02% to 3.81%, 1.35%, 2.81%, and 2.64% on Class B, C, D, and E, respectively. This is mainly because that the network without DMCP module only refines the pixels of the current block instead of compensating the block according to accurate motion fields estimated by DMCP module, which weakens the ability of the network to deal with the complex motions. However, the coding performance of the network without the DMCP module only reduces from 8.88% to 8.80% on *Cactus*. A possible reason is that the affine inter mode behaves well on *Cactus*, and it is difficult for the DMCP module to further improve the coding performance of the affine inter mode on *Cactus*.

The initial motion field I_{MF} is utilized to estimate accurate motion fields in the DMCP module. To evaluate its effectiveness, the I_{MF} is removed from the DMCP module. The

TABLE VI
COMPUTATIONAL COMPLEXITY OF THE PROPOSED METHOD

Class	Sequences	Inter DAMC-Net		Inter & Merge DAMC-Net	
		Enc	*Dec*	*Enc*	*Dec*
Class B	MarketPlace	274.23%	2283.18%	2329.43%	5601.85%
	RitualDance	255.95%	801.46%	1688.64%	2493.38%
	Cactus	220.15%	1031.24%	1780.65%	6383.36%
	BasketballDrive	245.58%	623.50%	1717.22%	2975.84%
	BQTerrace	186.56%	531.26%	2046.32%	2975.84%
	Average Class B	236.50%	1184.85%	1912.45%	4363.61%
Class C	RaceHorces	196.28%	681.36%	1412.65%	2693.71%
	BQMall	220.14%	413.58%	1962.21%	3219.39%
	PartyScene	201.40%	496.33%	1596.83%	3178.89%
	BasketballDrill	230.53%	534.94%	1931.33%	1922.85%
	Average Class C	212.09%	533.80%	1725.76%	2753.71%
Class D	RaceHorses	215.45%	538.58%	1507.74%	2409.37%
	BQSquare	187.88%	548.44%	1641.80%	3332.21%
	BlowingBubbles	204.16%	731.74%	1551.74%	2917.28%
	BasketballPass	204.16%	731.74%	1563.99%	1750.04%
	Average Class D	212.52%	548.06%	1566.32%	2605.90%
Class E	FourPeople	295.58%	725.78%	2277.04%	3502.82%
	Johnny	364.98%	1100.08%	2946.95%	2917.28%
	KristenAndSara	341.90%	1042.55%	2289.43%	3333.29%
	Average Class E	334.15%	956.14%	2504.47%	3251.13%
Overall		**242.71%**	**779.16%**	**1890.25%**	**3396.81%**

network without the I_{MF} achieves 2.80% BD-rate reduction on average, which is less than that of the proposed DAMC-Net. The experimental results prove that the initial motion field is helpful to improve the quality of learned motion fields. Moreover, in order to verify the benefit of the attention in the AFR module, the CBAM module is removed from the DAMC-Net. The coding performance of the network without attention decreases from 3.11% to 2.88% on average, which proves that the attention in the proposed DAMC-Net is useful to reconstruct the final output block.

E. Discussion

1) Complexity Analysis: It is widely known that the learning-based tools achieve superior coding efficiency at the cost of high computational complexity [15], [18]. To address this issue, in this paper, the network inference is running in GPU and the codec is performed in CPU. The encoding and decoding time of the proposed method in comparison with VTM-6.2 are shown in Table VI. Since the proposed DAMC-Net is integrated into the decision process of the optimal mode, the encoding time increases to 242.71% and 1890.25% for "Inter DAMC-Net" and "Inter & Merge DAMC-Net", respectively. The decoding time increases to 779.16% and 3,396.81%. Since the high complexity mainly comes from the forward operation in the network, the fluctuation of the decoding times is primarily determined by the number and size of CUs selected with the proposed DAMC mode.

TABLE VII
BD-RATE RESULTS OF THE PROPOSED METHOD ON AFFINE TEST SEQUENCES

Class	Sequences	Inter & Merge AMC	Inter & Merge DAMC-Net
Affine Test Sequences	BlueSky	-9.08%	-10.34%
	RollerCoaster	-4.87%	-6.71%
	Shields	-1.55%	-4.00%
	Cactus	-9.16%	-10.83%
Overall		**-6.17%**	**-7.97%**

TABLE VIII
BD-RATE RESULTS OF THE PROPOSED METHOD FOR VTM-12.1

Class	Inter & Merge AMC	Inter & Merge DAMC-Net
Class B	-4.24%	-5.70%
Class C	-1.73%	-3.35%
Class D	-3.26%	-5.88%
Class E	-3.24%	-4.57%
Overall	**-3.18%**	**-4.95%**

Furthermore, the storage consumption of the CNN models and the run-time GPU memory are also analyzed. There are 12 models for the proposed "Inter DAMC-Net" and 19 models for the proposed "Inter & Merge DAMC-Net". Specifically, the model size of the DAMC-Net is 4.36 MB. Hence, the total model size equals 52.52 MB and 82.84 MB for each QP, respectively. As for GPU memory usage, 1,762 MB and 1,968 MB run-time GPU memory are needed, respectively. In the future, we will explore more lightweight networks to better achieve this task.

2) Coding Performance on Affine Test Sequences: The coding performances on some affine test sequences [3] are reported to valid the efficacy of the proposed method to deal with complex motions. The results are shown in Table VII. Obviously, the proposed method achieves better coding performance than the traditional AMC. Moreover, the proposed "Inter & Merge DAMC-Net" achieves 7.97% BD-rate reduction on average, which is more than the average coding gain on CTC test sequences. It convinces that the proposed method is more suitable for sequences with complex motions.

3) Coding Performance for VTM-12.1: To further evaluate the effectiveness of the proposed method on the new reference software of VVC, the VTM-12.1 is used as the reference codec and the proposed DAMC-Net is incorporated into VTM-12.1 accordingly. The results are provided in Table VIII. Overall, the proposed "Inter & Merge DAMC-Net" achieves 4.95% BD-rate reduction on Y component. Furthermore, the overall R-D performance of the proposed method increases from

4.32% to 4.95% when compared with the results on VTM-6.2. Therefore, the proposed method is proved effective for the new reference software.

V. Conclusion

In this paper, a DAMC-Net for inter prediction is proposed to effectively boost the performance of affine motion compensation in VVC. In order to compensate the current encoding block, a DMCP module is designed to estimate accurate motion fields from spatial neighboring information, temporal reference block, and initial motion field. Then, an attention-based fusion and reconstruction module is designed to fuse multi-channel features from DMCP and reconstruct the final prediction signal. The proposed DAMC-Net is integrated into VVC as an optional mode for CU with affine mode. Experimental results demonstrate that the proposed DAMC-Net can considerably enhance coding performance.

References

[1] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1649–1668, Dec. 2012.

[2] S. Li, C. Zhu, and M.-T. Sun, "Hole filling with multiple reference views in DIBR view synthesis," *IEEE Trans. Multimedia*, vol. 20, no. 8, pp. 1948–1959, Aug. 2018.

[3] L. Li *et al.*, "An efficient four-parameter affine motion model for video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 8, pp. 1934–1948, Aug. 2018.

[4] J. Xie, N. He, L. Fang, and P. Ghamisi, "Multiscale densely-connected fusion networks for hyperspectral images classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 1, pp. 246–259, Jan. 2021.

[5] J. Lei *et al.*, "Deep stereoscopic image super-resolution via interaction module," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 8, pp. 3051–3061, Aug. 2021.

[6] L. Wang *et al.*, "Learning parallax attention for stereo image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12242–12251.

[7] Y. Fang, C. Zhang, H. Huang, and J. Lei, "Visual attention prediction for stereoscopic video by multi-module fully convolutional network," *IEEE Trans. Image Process.*, vol. 28, no. 11, pp. 5253–5265, Nov. 2019.

[8] W. Bao, W.-S. Lai, X. Zhang, Z. Gao, and M.-H. Yang, "MEMC-Net: Motion estimation and motion compensation driven neural network for video interpolation and enhancement," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 3, pp. 933–948, Mar. 2021.

[9] S. Huo, D. Liu, F. Wu, and H. Li, "Convolutional neural network-based motion compensation refinement for video coding," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, May 2018, pp. 1–4.

[10] 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 Trans. Circuits Syst. Video Technol.*, vol. 29, no. 11, pp. 3291–3301, Nov. 2019.

[11] J. Mao and L. Yu, "Convolutional neural network based bi-prediction utilizing spatial and temporal information in video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 7, pp. 1856–1870, Jul. 2020.

[12] Y. Wang, X. Fan, C. Jia, D. Zhao, and W. Gao, "Neural network based inter prediction for HEVC," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2018, pp. 1–4.

[13] N. Yan, D. Liu, H. Li, B. Li, L. Li, and F. Wu, "Convolutional neural network-based fractional-pixel motion compensation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 3, pp. 840–853, Mar. 2019.

[14] J. Lin, D. Liu, H. Li, and F. Wu, "Generative adversarial network-based frame extrapolation for video coding," in *Proc. IEEE Vis. Commun. Image Process. (VCIP)*, Dec. 2018, pp. 1–4.

[15] S. Huo, D. Liu, B. Li, S. Ma, F. Wu, and W. Gao, "Deep network-based frame extrapolation with reference frame alignment," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 3, pp. 1178–1192, Mar. 2021.

[16] L. Zhao, S. Wang, X. Zhang, S. Wang, S. Ma, and W. Gao, "Enhanced motion-compensated video coding with deep virtual reference frame generation," *IEEE Trans. Image Process.*, vol. 28, no. 10, pp. 4832–4844, Oct. 2019.

[17] S. Xia, W. Yang, Y. Hu, and J. Liu, "Deep inter prediction via pixel-wise motion oriented reference generation," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 1710–1714.

[18] J. Liu, S. Xia, and W. Yang, "Deep reference generation with multi-domain hierarchical constraints for inter prediction," *IEEE Trans. Multimedia*, vol. 22, no. 10, pp. 2497–2510, Oct. 2020.

[19] H. Choi and I. V. Bajić, "Deep frame prediction for video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 7, pp. 1843–1855, Jul. 2020.

[20] H. Choi and I. V. Bajic, "Affine transformation-based deep frame prediction," *IEEE Trans. Image Process.*, vol. 30, pp. 3321–3334, 2021.

[21] J. Lin, D. Liu, H. Yang, H. Li, and F. Wu, "Convolutional neural network-based block up-sampling for HEVC," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 12, pp. 3701–3715, Dec. 2019.

[22] Y. Li *et al.*, "Convolutional neural network-based block up-sampling for intra frame coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 9, pp. 2316–2330, Sep. 2018.

[23] J. Deng, L. Wang, S. Pu, and C. Zhuo, "Spatio-temporal deformable convolution for compressed video quality enhancement," in *Proc. AAAI*, Apr. 2020, pp. 10696–10703.

[24] Z. Pan, X. Yi, Y. Zhang, B. Jeon, and S. Kwong, "Efficient in-loop filtering based on enhanced deep convolutional neural networks for HEVC," *IEEE Trans. Image Process.*, vol. 29, pp. 5352–5366, 2020.

[25] Z. Guan, Q. Xing, M. Xu, R. Yang, T. Liu, and Z. Wang, "MFQE 2.0: A new approach for multi-frame quality enhancement on compressed video," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 3, pp. 949–963, Mar. 2021.

[26] S. Ma, X. Zhang, C. Jia, Z. Zhao, S. Wang, and S. Wang, "Image and video compression with neural networks: A review," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 6, pp. 1683–1698, Jun. 2020.

[27] J. Lin, D. Liu, H. Li, and F. Wu, "M-LVC: Multiple frames prediction for learned video compression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3543–3551.

[28] E. Agustsson, D. Minnen, N. Johnston, J. Balle, S. J. Hwang, and G. Toderici, "Scale-space flow for end-to-end optimized video compression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8500–8509.

[29] R. Yang, F. Mentzer, L. Van Gool, and R. Timofte, "Learning for video compression with hierarchical quality and recurrent enhancement," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6627–6636.

[30] Z. Chen, T. He, X. Jin, and F. Wu, "Learning for video compression," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 2, pp. 566–576, Feb. 2020.

[31] G. Lu, W. Ouyang, D. Xu, X. Zhang, C. Cai, and Z. Gao, "DVC: An end-to-end deep video compression framework," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10998–11007.

[32] K. Zhang, Y.-W. Chen, L. Zhang, W.-J. Chien, and M. Karczewicz, "An improved framework of affine motion compensation in video coding," *IEEE Trans. Image Process.*, vol. 28, no. 3, pp. 1456–1469, Mar. 2019.

[33] H. Huang, J. W. Woods, Y. Zhao, and H. Bai, "Control-point representation and differential coding affine-motion compensation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 10, pp. 1651–1660, Oct. 2013.

[34] J. Dai *et al.*, "Deformable convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 764–773.

[35] S. Woo, J. Park, J. Lee, and I. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Jul. 2018, pp. 3–19.

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

[37] Xiph.org. (2017). *Xiph.org Video Test Media*. [Online]. Available: https://media.xiph.org/video/derf

[38] VQEG. (2017). *VQEG Video Datasets Organizations*. [Online]. Available: https://www.its.bldrdoc.gov/vqeg/video-datasetsand-organizations.aspx/

[39] L. Song, X. Tang, W. Zhang, X. Yang, and P. Xia, "The SJTU 4K video sequence dataset," in *Proc. 5th Int. Workshop Qual. Multimedia Exper. (QoMEX)*, Jul. 2013, pp. 34–35.

[40] K. Suehring and X. Li, *JVET Common Test Conditions and Software Reference Configurations*, document JVET-G1010, Aug. 2017.

[41] J. Lei, J. Duan, W. Feng, N. Ling, and C. Hou, "Fast mode decision based on grayscale similarity and inter-view correlation for depth map coding in 3D-HEVC," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 3, pp. 706–718, Mar. 2018.

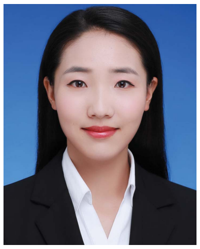
[42] G. Bjontegaard, *Calculation of Average PSNR Differences Between RD Curves*, document VCEG-M33, Apr. 2001.

[43] M. Abadi *et al.*, "TensorFlow: Large-scale machine learning on heterogeneous distributed systems," 2016, *arXiv:1603.04467*. [Online]. Available: http://arxiv.org/abs/1603.04467

[44] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980

Dengchao Jin received the B.S. degree in mechatronic engineering from Northwestern Polytechnical University, Xi'an, China, in 2019. He is currently pursuing the Ph.D. degree with the School of Electrical and Information Engineering, Tianjin University, Tianjin, China. His research interests include video coding and deep learning.

Jianjun Lei (Senior Member, IEEE) received the Ph.D. degree in signal and information processing from Beijing University of Posts and Telecommunications, Beijing, China, in 2007. He was a Visiting Researcher with the Department of Electrical Engineering, University of Washington, Seattle, WA, USA, from August 2012 to August 2013. He is currently a Professor with Tianjin University, Tianjin, China. His research interests include 3-D video processing, virtual reality, and artificial intelligence.



Bo Peng (Member, IEEE) received the Ph.D. degree in information and communication engineering from Tianjin University, Tianjin, China, in 2020. She was a Visiting Research Scholar with the School of Computing, National University of Singapore, Singapore, from March 2019 to April 2020. She is currently an Assistant Professor with the School of Electrical and Information Engineering, Tianjin University. Her research interests include computer vision, image processing, and vision understanding.

Wanqing Li (Senior Member, IEEE) received the Ph.D. degree in electronic engineering from The University of Western Australia. He was an Associate Professor with Zhejiang University from 1991 to 1992, a Senior Researcher and later a Principal Researcher with Motorola Research Laboratory from 1998 to 2003, and a Visiting Researcher with Microsoft Research, USA, in 2008, 2010, and 2013. He is currently an Associate Professor and the Director of the Advanced Multimedia Research Laboratory (AMRL), University of Wollongong, Australia. His research areas include machine learning, 3D computer vision, 3D multimedia signal processing, and medical image analysis. He also serves as an Associate Editor for IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY and IEEE TRANSACTIONS ON MULTIMEDIA.

Nam Ling (Life Fellow, IEEE) received the B.Eng. degree from the National University of Singapore, Singapore, in 1981, and the M.S. and Ph.D. degrees from the University of Louisiana at Lafayette, Lafayette, LA, USA, in 1985 and 1989, respectively. From 2002 to 2010, he was an Associate Dean with the School of Engineering, Santa Clara University, Santa Clara, CA, USA. He was Sanfilippo Family Chair Professor, and currently Wilmot J. Nicholson Family Chair Professor and the Chair with the Department of Computer Science and Engineering, Santa Clara University. He is/was also a Consulting Professor with the National University of Singapore; a Guest Professor with Tianjin University, Tianjin, China, and Shanghai Jiao Tong University, Shanghai, China; Cuiying Chair Professor with Lanzhou University, Lanzhou, China; a Chair Professor and Minjiang Scholar with Fuzhou University, Fuzhou, China; a Distinguished Professor with Xi'an University of Posts and Telecommunications, Xi'an, China; a Guest Professor with Zhongyuan University of Technology, Zhengzhou, China; and an Outstanding Overseas Scholar with Shanghai University of Electric Power, Shanghai. He has authored or coauthored over 220 publications and seven adopted standard contributions. He has been granted nearly 20 U.S. patents so far. He is an IEEE Fellow due to his contributions to video coding algorithms and architectures. He is also an IET Fellow. He was named as an IEEE Distinguished Lecturer twice and also an APSIPA Distinguished Lecturer. He was a recipient of the IEEE ICCE Best Paper Award (First Place) and the Umedia Best/Excellent Paper Award three times, six awards from Santa Clara University, four at the University level (Outstanding Achievement, Recent Achievement in Scholarship, President's Recognition, and Sustained Excellence in Scholarship), and two at the School/College level (Researcher of the Year and Teaching Excellence). He was a Keynote Speaker of IEEE APCCAS, VCVP (twice), JCPC, IEEE ICAST, IEEE ICIEA, IET FC Umedia, IEEE Umedia, IEEE ICCIT, and Workshop at XUPT (twice); and a Distinguished Speaker of IEEE ICIEA. He served as the General Chair/Co-Chair for IEEE Hot Chips, VCVP (twice), IEEE ICME, Umedia (seven times), and IEEE SiPS. He was an Honorary Co-Chair of IEEE Umedia 2017. He served as the Technical Program Co-Chair for IEEE ISCAS, APSIPA ASC, IEEE APCCAS, IEEE SiPS (twice), DCV, and IEEE VCIP. He was the Technical Committee Chair of IEEE CASCOM TC and IEEE TCMM. He served as a Guest Editor or an Associate Editor for the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS—I: REGULAR PAPERS, the IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING, IEEE ACCESS, JSPS (Springer), and MSSP (Springer).

Qingming Huang (Fellow, IEEE) received the bachelor's degree in computer science and the Ph.D. degree in computer engineering from Harbin Institute of Technology, China, in 1988 and 1994, respectively. He is currently a Professor with the University of Chinese Academy of Sciences and an Adjunct Research Professor with the Institute of Computing Technology, Chinese Academy of Sciences. He has published more than 400 academic articles in prestigious international journals, including the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON MULTIMEDIA, and the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, and top-level conferences, such as the ACM Multimedia, ICCV, CVPR, IJCAI, and VLDB. His research areas include multimedia video analysis, image processing, computer vision, and pattern recognition. He served as the General Chair, the Program Chair, the Track Chair, and a TPC Member for various conferences, including ACM Multimedia, CVPR, ICCV, ICME, PCM, and PSIVT. He is also an Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY and *Acta Automatica Sinica*, and a Reviewer of various international journals, including the IEEE TRANSACTIONS ON MULTIMEDIA, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, and the IEEE TRANSACTIONS ON IMAGE PROCESSING.