

[AHG11 & AHG6] DOVC: Deep Omnidirectional Video Compression

Qipu Qin¹, Cheolkon Jung¹, Dan Zou², and Ming Li²

¹Xidian University, China

²OPPO, China

Introduction

2



■ Background

- Deep learning-based video compression (DLVC or DVC) has achieved great advances in improving coding efficiency
- DLVC/DVC can be summarized into the following two aspects:
 - (1) Combine deep learning with traditional hybrid video compression (Module-level)
 - (2) **Establish a novel deep video compression framework** (Codec-level)

■ Motivation

- Compression of omnidirectional videos (6K, 8K) using NN
- High complexity and module-level local optimization
- Limit of the optical flow network



Proposed Solution

3

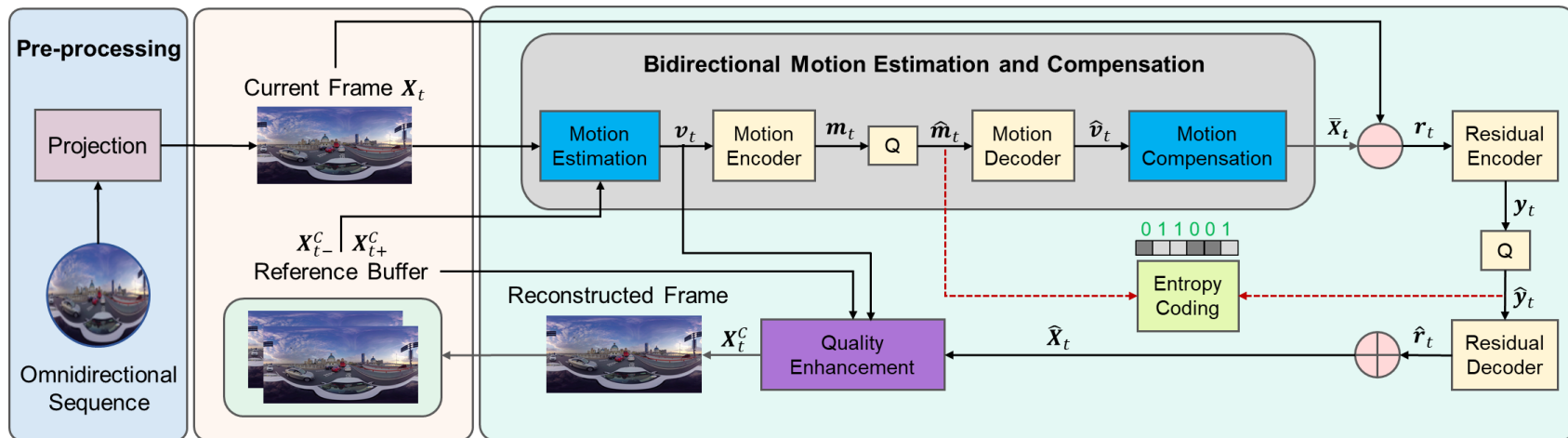
■ Network Architecture (Encoder)

DOVC mainly contains:

- Projection
- Bidirectional motion estimation
- Motion encoder/decoder

End-to-end deep omnidirectional video compression network (DOVC)

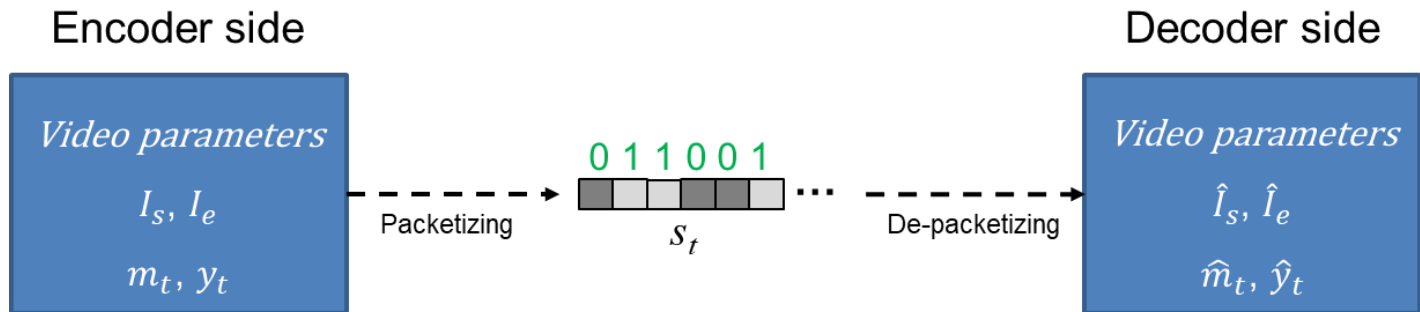
- Bidirectional motion compensation
- Residual encoder/decoder
- Quality enhancement
- Entropy coding



Proposed Solution

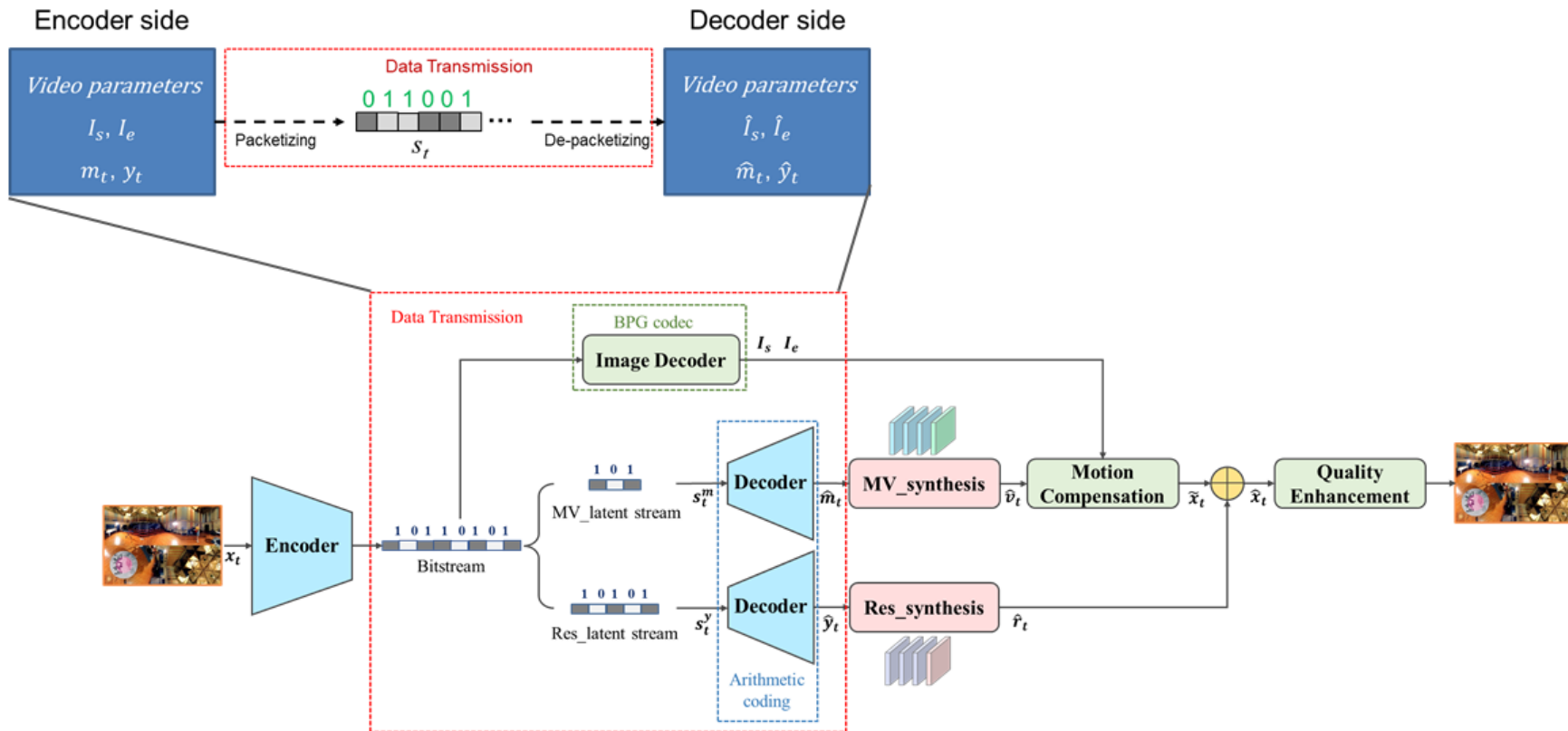
■ Entropy Coding for Bitstream

- For I frame, DOVC encoder uses **BPG tool for image compression** and decompression
- For B/P frame, DOVC encoder encodes **motion and residual** after quantization
- **Entropy coding** encodes feature maps into bitstream and decodes bitstream into feature maps



Proposed Solution

■ Network Architecture (Decoder)



■ Inference Stage

Network Information in Inference Stage		
Mandatory	GPU Type	GeForce GTX1080 Ti
	Framework:	Python3.6, PyTorch1.6, CUDA10.1
	Number of GPUs per Task	1
	Total Parameter Number	19.1 M for $\lambda = 256$ model, 19.1M for $\lambda = 512$ model, 19.1 M for $\lambda = 1024$ model, 19.1M for $\lambda = 2048$ model
	Parameter Precision (Bits)	32 (F)
	Memory Parameter (MB)	122.54 MB for each model
	Test data information:	360-degree sequences [3] provided by JVET
	MAC (Giga)	$kernel^2 \times \sum (in_channels \times out_channels)$
Optional	Total Conv. Layers	DOVC is composed of a series of CNNs, including a total of 6 modules.
	Total FC Layers	0
	Total Memory (MB)	/
	Batch size:	4
	Patch size	Whole frame
	Changes to network configuration or weights required to generate rate points	/
	Peak Memory Usage	/
	Other information:	/

■ Training Stage

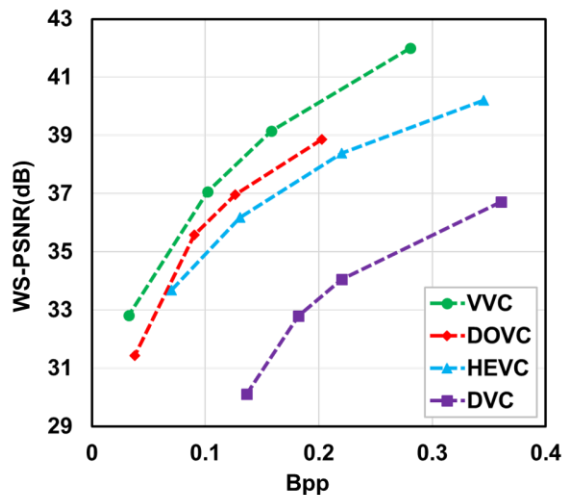
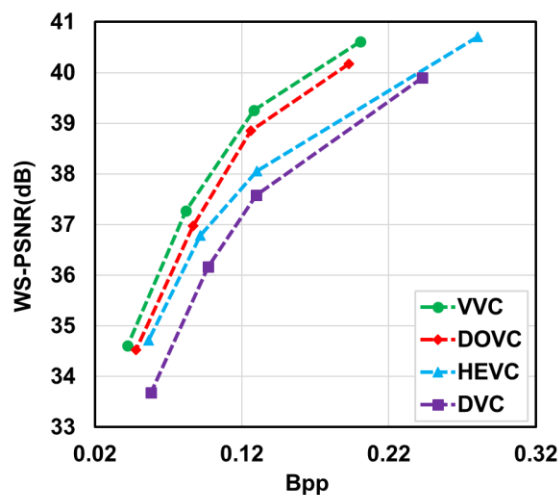
Network Information in Training Stage		
Mandatory	GPU Type	Nvidia Tesla V100, 32G
	Number of GPUs per Task	2
	Framework:	Python3.6, PyTorch1.6, CUDA10.0
	Epoch:	100
	Batch size:	4
	Loss function:	RDO: $R + \lambda \cdot WMSE$ Note that R represents the total number of bits for encoding, λ is Lagrangian multiplier, and $WMSE$ is a weighted MSE based on different sphere-to-plane projection formats.
	Training time:	About 96 hours / model
	Training data information:	VQA-ODV dataset [9]
Optional	Training configurations for generating compressed training data (if different to VTM CTC):	$\lambda = 256, 512, 1024, 2048$ DOVC
	Number of iterations:	2000000
	Patch size:	1280×1280
	Learning rate:	1e-4
	Optimizer:	ADAM
	Preprocessing:	See description in main text
	Mini-batch selection process:	/
	Other information:	/
	Preprocessing:	Projection: ERP and CMP

■ Experimental Setup

- Training environment: Nvidia Tesla V100 (32G), Python3.6, PyTorch1.6, CUDA10.0
- Test environment: GeForce GTX1080Ti (12G), Python3.6, PyTorch1.6, CUDA10.1
- **Anchors:** Intel Xeon CPU (Dual processor, RAM 32.G), HM-16.16 (with 360Lib-5.0), VTM-11.0 (with 360Lib-12.0), Visual Studio 2013
- **Training dataset:** VQA-ODV^[1]
- **Test dataset:** 360-degree sequences provided by JVET
- Other details: Epoch = 100, Batch Size = 4, Learning Rate = 1e-4, Patch Size = 1280*1280, Optimizer = ADAM, $\lambda = 256, 512, 1024, 2048$

^[1] Proc. ACM Multimedia 2018

Objective Comparison



DOVC, DVC: RGB domain (4:4:4)
VVC, HEVC: YUV domain (4:2:0)

Rate-distortion performance in terms of WS-PSNR.

BD-BR (%) and BD-WS_PSNR (dB) performances of DOVC method in comparison with HEVC/H.265 and DVC

	DOVC vs HEVC/H.265		DOVC vs DVC		DOVC vs VVC/H.266	
Projection Format	CMP	ERP	CMP	ERP	CMP	ERP
BD-BR (%)	-14.2889	-27.7762	-27.7760	-58.8419	12.8983	38.9589
BD-WS_PSNR (dB)	0.4837	0.7597	1.2531	3.9060	-0.4705	-1.3186

■ Encoding Time Comparison

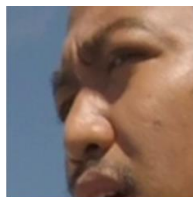
Comparison of coding time complexity of HEVC, VVC and DOVC

Class	Sequences	Coding Time Complexity Comparison (s)		
		VVC	HEVC	DOVC
S1	ChairliftRide (10bit)	238798.367	84503.725	1512.226
	Gaslamp	70911.438	32773.284	1475.378
	Harbor	133933.894	48565.107	1522.904
	KiteFlite	363812.892	104027.615	1506.412
	SkateboardInLot (10bit)	183657.670	65565.102	1460.329
	Trolley	90323.569	41697.319	1494.871
S2	Balboa	629755.557	112035.185	1456.837
	BranCastle2	162387.403	52067.384	1488.358
	Broadway	639893.832	124027.615	3117.108
	Landings2	89323.569	41565.107	1468.843
Average		260279.819	70682.744	1650.327

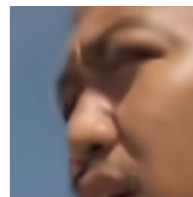
Visual Comparison



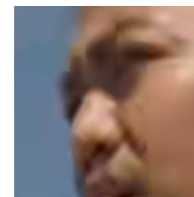
(a) SkateboardInLot (frame = 158)



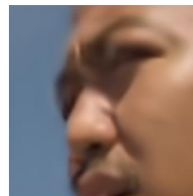
(b) Raw



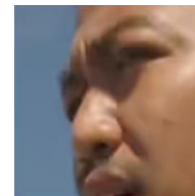
(c) DOVC
0.07450 bpp, 37.02dB



(d) DVC
0.07975 bpp, 36.68dB



(e) HEVC/H.265
0.07621 bpp, 37.08dB



(f) VVC/H.266
0.07425 bpp, 37.64dB

Visual comparison in SkateboardInLot (CMP format)

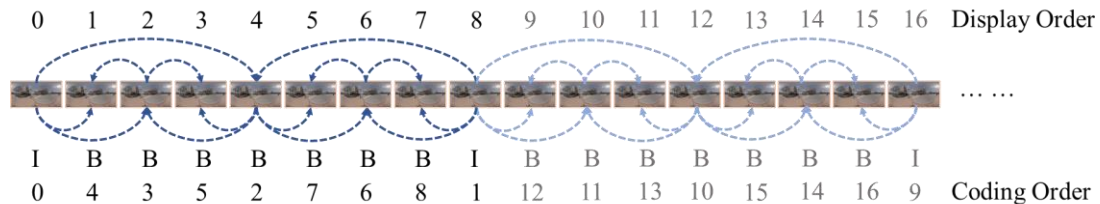
- **Contribution:** An end-to-end deep omnidirectional video compression framework (DOVC) with CNNs.
- **BD-BR and BD-WS_PSNR:** DOVC achieves average 21% reduction in BD-BR and average 0.6217dB gain in BD-WS_PSNR over HM-16.16 (360Lib-5.0) under LDP configuration for encoding omnidirectional videos.
- **Coding time of DOVC:** only 0.0234 times that of HM-16.16 (360Lib-5.0) and 0.0064 times that of VTM-11.0 (360Lib-12.0).

■ Recommendation to JVET:

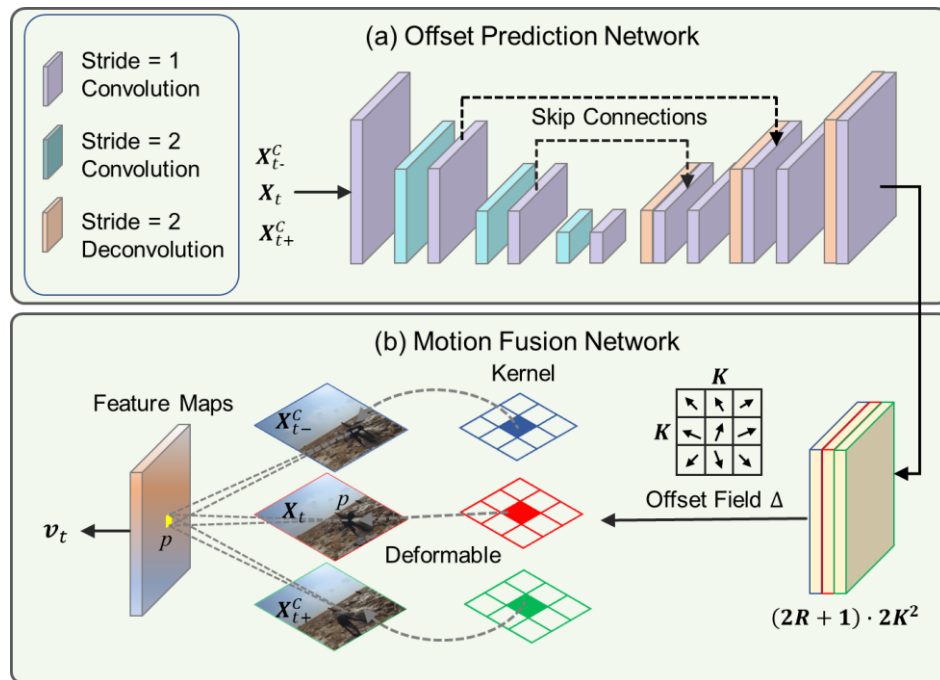
- For omnidirectional videos with much higher resolution (6K and 8K), it is necessary to explore an efficient compression framework for them.
- Deep learning shows outstanding non-linear fitting ability, which can be successfully applied to the omnidirectional video compression.
- Therefore, we propose a new **EE on coding omnidirectional videos** using deep NNs or including this topic in an existing EE. We recommend JVET to consider investigating this topic for further research.

Appendix

■ Bidirectional Motion Estimation

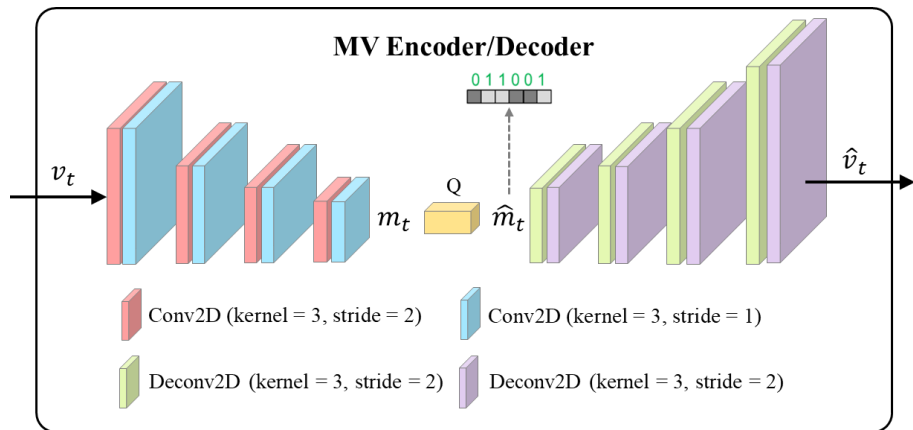


- Offset prediction network
- Offset field Δ
- **Motion fusion: Deformable convolution**
- Motion vector

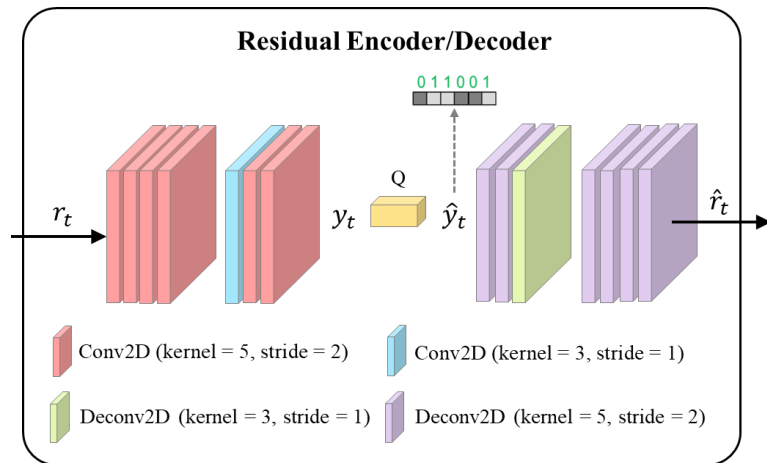


■ Motion and Residual Encoder/Decoder

- Auto-encoder style network to encode/decode motion vector and residual information
- Quantization
- \hat{m}_t and \hat{y}_t are sent to the entropy coding module for generating bitstream



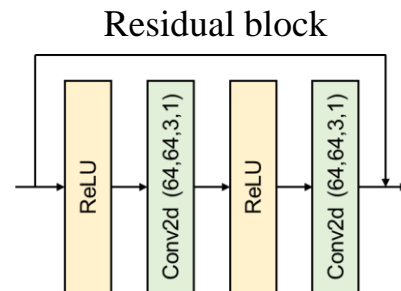
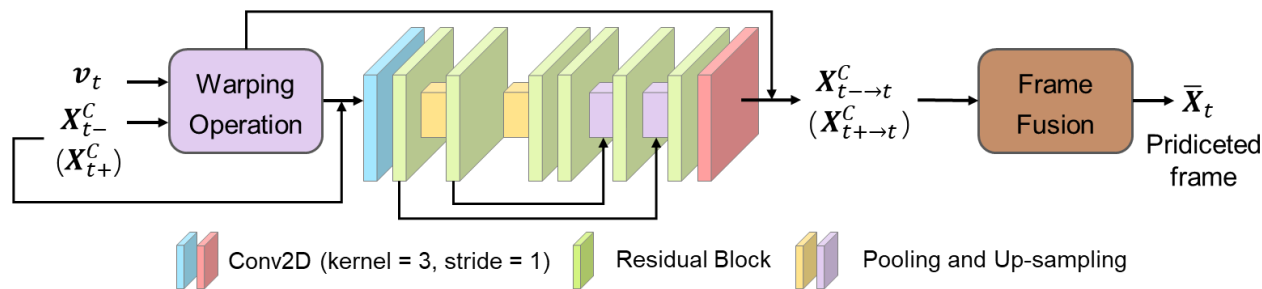
Motion encoder/decoder



Residual encoder/decoder

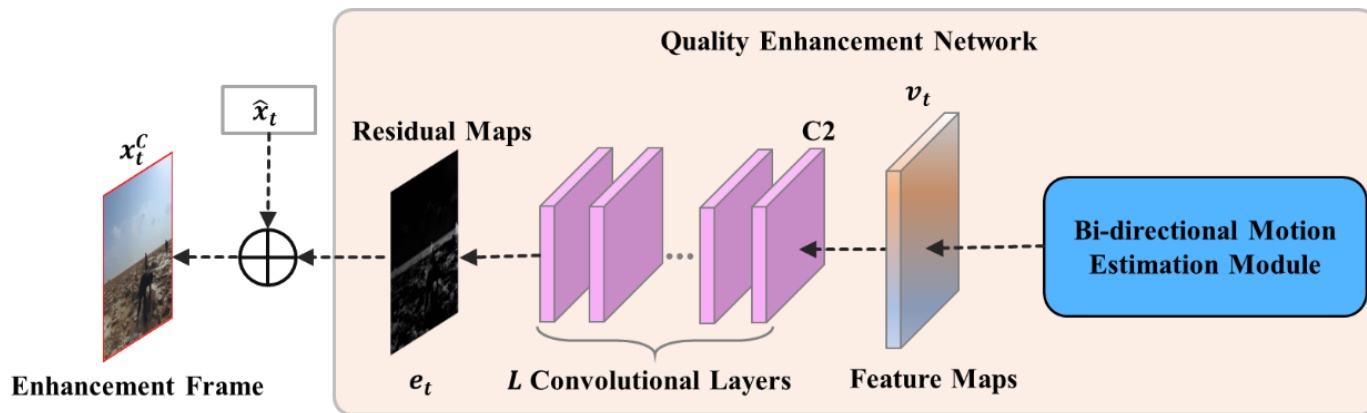
■ Bidirectional Motion Compensation

- Bidirectional prediction mode
- Warping operation
- Convolutional layers (2) + Residual blocks (6)
- Frame fusion



■ Quality Enhancement

- Reuse temporal information from motion estimation
- Output enhanced residual map
- Simple but effective



■ Projection and Loss Function

- Pre-processing stage: Sphere-to-plane projection
- Most popular formats: ERP and CMP
- Weighted factors of ERP and CMP
- Loss Function:

$$w_{erp}(i, j) = \cos \left(\left(j - \frac{Height}{2} + \frac{1}{2} \right) \cdot \frac{\pi}{Height} \right)$$

$$w_{cmp}(i, j) = \left(3 + \frac{(i+1)^2 + (j+1)^2 - (i+j) \cdot a}{a^2/4} \right)^{-3/2}$$

$$W(i, j) = \frac{w(i, j)}{\sum_{i=0}^{Width-1} \sum_{j=0}^{Height-1} w(i, j)}$$

$$WMSE = \sum_{i=0}^{Width-1} \sum_{j=0}^{Height-1} \left(\mathbf{X}(i, j) - \hat{\mathbf{X}}'(i, j) \right)^2 \cdot W(i, j)$$

$$\min \left\{ J = \mathbf{R} + \lambda \cdot WMSE \left(\mathbf{X}_t, \hat{\mathbf{X}}_t \right) \right\} \quad \text{Loss Function}$$

■ Ablation Study

- Analysis of Motion Estimation and Compensation Modules



(a) 5th Frame (Raw)



(b) Fused Offset map (DOVC)



(c) Optical Flow map



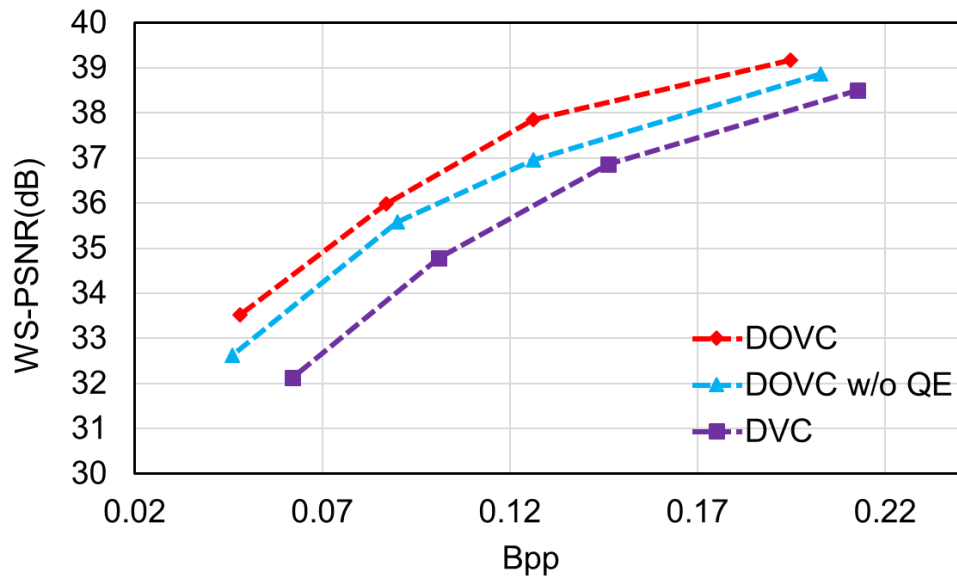
(d) Bidirectional compensation
(DOVC): 32.5dB



(e) Unidirectional compensation
(Optical Flow): 31.7dB

Visualization results on SkateboardInLot (CMP format) of ablation studies for motion estimation and motion compensation.

- Analysis of Quality Enhancement Module



Ablation study on the quality enhancement in DOVC.

The quality enhancement module is removed in DOVC. The results of DOVC without QE drop nearly 0.23dB when compared with the complete DOVC model. Although the quality enhancement module is removed, the performance of DOVC without QE still surpasses DVC.