Learned Video Compression with Spatial-Temporal Optimization

Yiming Wang, Qian Huang, Bin Tang, Wenting Liu, Wenchao Shan, Qian Xu
College of Computer Science and Software Engineering, Hohai University, Nanjing, China

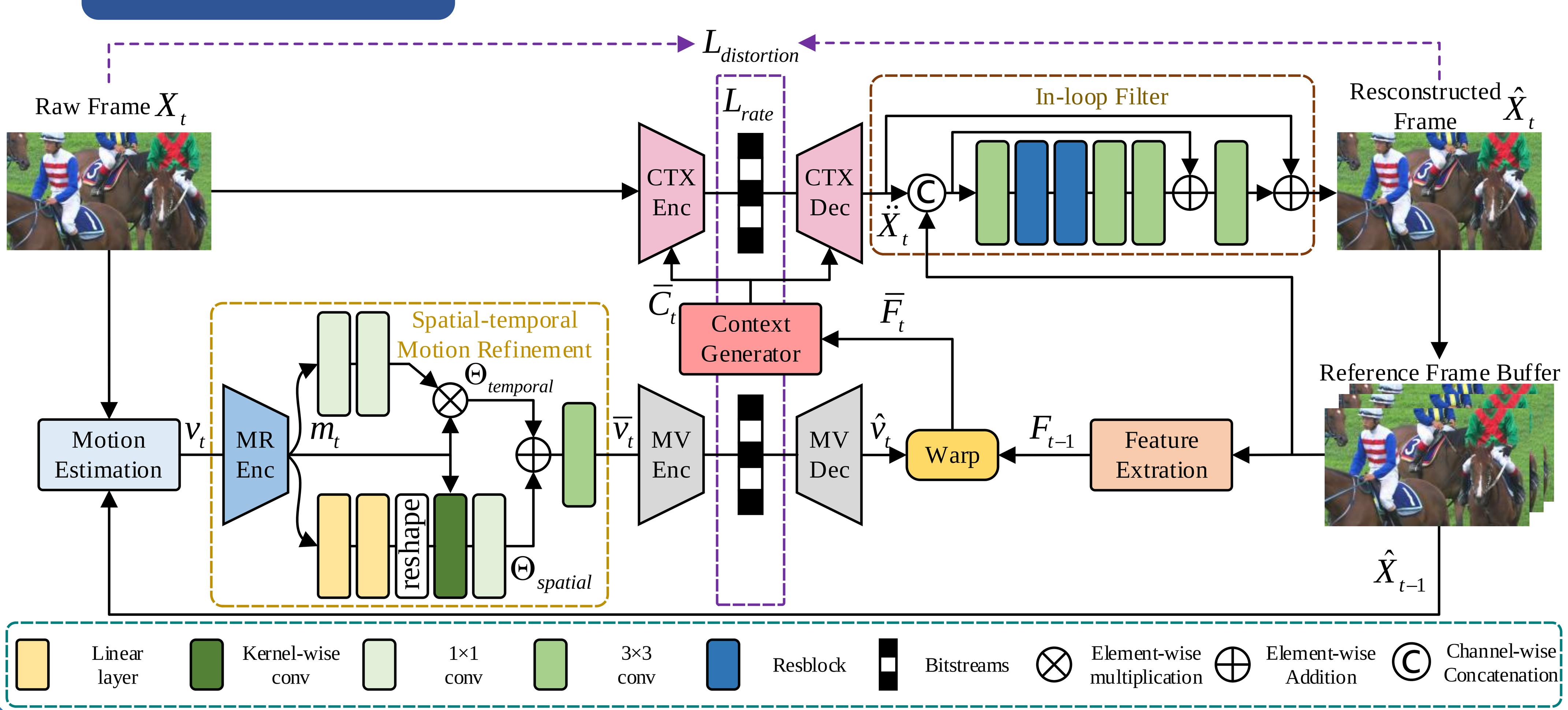
ICASSP 2024 KOREA

Background

Motivation: optical flow network is not accurate and may introduce extra artifacts.

- **Pixel-level optical flow based schemes:** it replaces the block motion estimation in traditional video coding with pre-trained optical flow model to estimate motion information. Inaccurate optical flow estimation may introduce the reconstructed artifacts.
- **Feature-level DCN based schemes:** it utilizes the DCN to extract motion information by performing feature alignment in an unsupervised manner. It is difficult to train in practice and results in offset maps overflow for lack of explicit guidance.

Methods



- We first propose a **spatial-temporal motion refinement (STMR)** module to extract spatial and temporal components to enhance the original MV for prediction.
- We adopt the popular **context coding scheme** instead of the residual coding scheme, mainly because $H(X_t - \bar{X}_t) \geq H(X_t \mid \bar{X}_t)$, H denotes the Shannon entropy, X_t and $\bar{X}_t$ denote the current and predicted frame, respectively.
- We propose an **in-loop filter (ILF)** module, which removes compression artifacts.
- The experiments demonstrate the coding performance of our proposed method.

Results

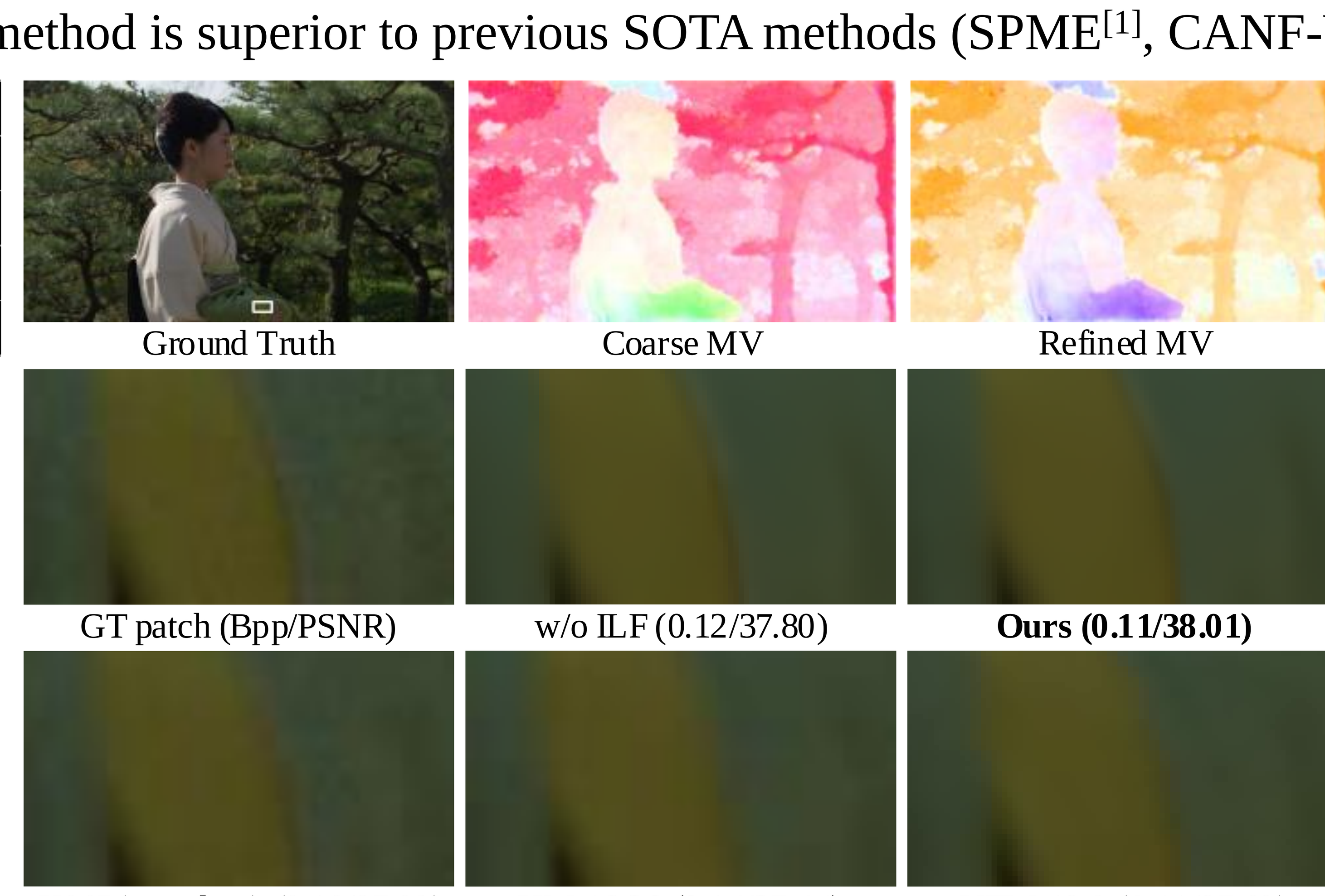
Sequences	PSNR							MS-SSIM						
	HM-16.20	VTM-13.2	SPME (DCVC)	CANF-VC	DMVC	HDCVC	Ours	HM-16.20	VTM-13.2	SPME (DCVC)	CANF-VC	DMVC	HDCVC	Ours
HEVC Class B	-30.53	-53.23	-35.56	-33.15	-32.12	-27.59	**-40.23**	-14.21	-41.24	-47.60	-48.02	-51.12	-48.24	**-63.79**
HEVC Class C	-18.59	-42.39	-10.72	-13.24	-11.34	-10.47	**-18.41**	-7.93	-34.18	-38.99	-44.76	-43.52	-43.07	**-52.81**
HEVC Class D	-17.50	-40.34	-15.48	-15.68	-18.51	-9.48	**-25.41**	-5.86	-31.82	-48.75	-50.64	-54.57	-49.70	**-56.63**
UVG	-30.26	-53.17	-48.75	-48.35	-36.68	-32.79	**-49.80**	-14.10	-41.80	-49.26	-50.64	-49.28	-45.98	**-58.45**
MCL-JCV	-17.57	-40.44	-30.30	-27.02	-14.81	-26.63	**-30.85**	-2.39	-32.36	-45.03	-47.26	-45.56	-49.77	**-57.16**
Average	-22.89	-45.91	-28.16	-27.49	-22.69	-21.39	**-32.94**	-8.90	-36.28	-45.93	-48.26	-48.81	-47.35	**-57.77**

- HM-16.20, VTM-13.2 — official reference software of H.265/HEVC, H.266/VVC.
- The best result of learned method is highlighted.

- In terms of **PSNR** metrics, our proposed method achieves comparable results with VTM-13.2 and even exceed it on 1080p dataset.
- In terms of **MS-SSIM** metrics, our method is superior to previous SOTA methods (SPME[1], CANF-VC[2], DMVC[3],HDCVC[4]) by a larger margin.

STMR	ILF	B	C	D
✔	✔	0.0	0.0	0.0
✖	✔	8.75	9.23	7.40
✔	✖	9.81	13.00	13.46
✖	✖	12.45	18.89	18.68

- Our proposed modules all save BD-rates to varying degrees.
- Right figure shows that refined MV has a richer structural texture.
- One patch after using our method is visually more like the ground truth (GT) patch while consuming fewer bits.



Scheme	Enc speed	Dec Speed
x265(veryslow)	0.25	19.2
HM-16.20	0.02	9.9
VTM-13.2	0.001	1.1
CANF-VC	0.6	0.9
DMVC	1.8	-
HDCVC	2.4	1.78
Ours	2.08	2.56

- Our method achieve the faster speed (2.56 FPS) than other methods.
- we are slower than HDCVC[4] because it removes the time-consuming multi-frame enhancement module in the test.

Conclusion

- We propose the learned video compression with spatial-temporal optimization. In particular, spatial-temporal motion refinement module is proposed to refine the MV.
- In-loop filter module is proposed to remove compression artifacts and finally enhance the reconstruction quality.
- Qualitative and quantitative experiments have shown that our method outperforms the recent learned methods in terms of both PSNR and MS-SSIM metrics.

Reference

[1] Han Gao, Jinzhong Cui, Mao Ye, Shuai Li, Yu Zhao, and Xiatian Zhu, "Structure-preserving motion estimation for learned video compression," in ACM MM, 2022, p.3055–3063.

[2] Yung-Han Ho, Chih-Peng Chang, Peng-Yu Chen, Alessandro Gnutti, and Wen-Hsiao Peng, "CANFVC: conditional augmented normalizing flows for video compression," in ECCV, 2022, pp. 207–223.

[3] Kai Lin, Chuanmin Jia, Xinfeng Zhang, Shanshe Wang, Siwei Ma, and Wen Gao, "DMVC: Decomposed motion modeling for learned video compression," IEEE TCSVT, pp. 3502–3515, 2023.

[4] HuairuiWang, Zhenzhong Chen, and ChangWen Chen, "Learned video compression via heterogeneous deformable compensation network," IEEE TMM, pp. 1–12, 2023.

Contact

Prof. Qian Huang
huangqian@hhu.edu.cn

This work is supported by Fundamental Research Funds of China for the Central Universities (B210201053,B230205048).