High Dynamic Range Video Compression: A Large-Scale Benchmark Dataset and A Learned Bit-depth Scalable Compression Algorithm

Supplementary Material

7. Dataset

More video clips of HDRVD2K are presented in Fig. 7. We can find that these clips contain multiple scenes and different luminance conditions.

8. Network Details

We present the specific structure of mv encoder-decoder in EL coding, which as shown in Fig. 8. The contextual MV encoder-decoder designed in [7] is utilized to encode and decode the HDR frame MV v_t^e for the EL. MV contexts C_t^v are extracted by an MV context extractor and fed into the contextual MV encoder-decoder symmetrically to reduce the interlayer motion redundancy. To further reduce the redundancy in HDR frame MV latent, we incorporate the MV hyper prior f_h^v and the MV contextual prior f_c^v derived from the MV contexts to estimate more accurately the probability distribution parameters of HR MV latent representations.

we also present the specific structure of hybrid temporal-layer context mining module, which as shown in Fig. 9. Temporal features and spatial features have different properties. To utilize the advantages of both, we use Hybrid Temporal-Layer Context Mining (HTLCM) designed in [7] to generate hybrid contexts for encoding the HDR frame. Instead of generating contexts only from HDR video temporal features $\hat{F}_{t-1}^e$, we use a weight map fusion module to integrate both the temporal features $\hat{F}_{t-1}^e$ and the spatial features from $\hat{F}_t^b$. After context mining, multiscale hybrid contexts C_t^1, C_t^2 and C_t^3 are fed into the contextual encoder-decoder [31] to compress HDR frames.

9. Encoding time

We supplement the encoding/decoding time of different methods, as shown in Table 6. On the one hand, it can be found that the encoding time of LVC methods significantly less than that of SHM. On the other hand, for LVC methods, the encoding time of our LBSVC is only second to that of HEM*, while achieving the optimal compression performance on HDR videos.

10. Base layer Performance Comparison

BL compression performance of different methods are shown in Fig. 10. We can find BL performance of our LBSVC is better than SHM [49] and LSSVC. The BL framework in LBSVC is HEM [31], which is also the BL of comparison method HEM*. Meanwhile, the BL framework of LSSVC [7] is TCM [52], whose performance is similar to SHM [49]. In fact, BL of SHM is HM, which is also the BL of Mai11 [38] in this paper. We can find that BL of our LBSVC outperforms other methods in JVET and HDM test dataset.

Table 6. Average encoding/decoding time for one 1080p frame. A NVIDIA RTX A6000 GPU and a Intel(R) Core(TM) i7-11700 CPU are corresponding platform for different methods.

Methods	Enc. time (s)		Dec. time (s)		Platform
	BL	EL	BL	EL	
SHM	32.82	36.84	0.05	0.06	CPU
HEM*	0.54	0.54	0.19	0.19	GPU
LSSVC	0.78	1.62	0.49	1.35	GPU
LBSVC	0.72	1.64	0.24	0.81	GPU

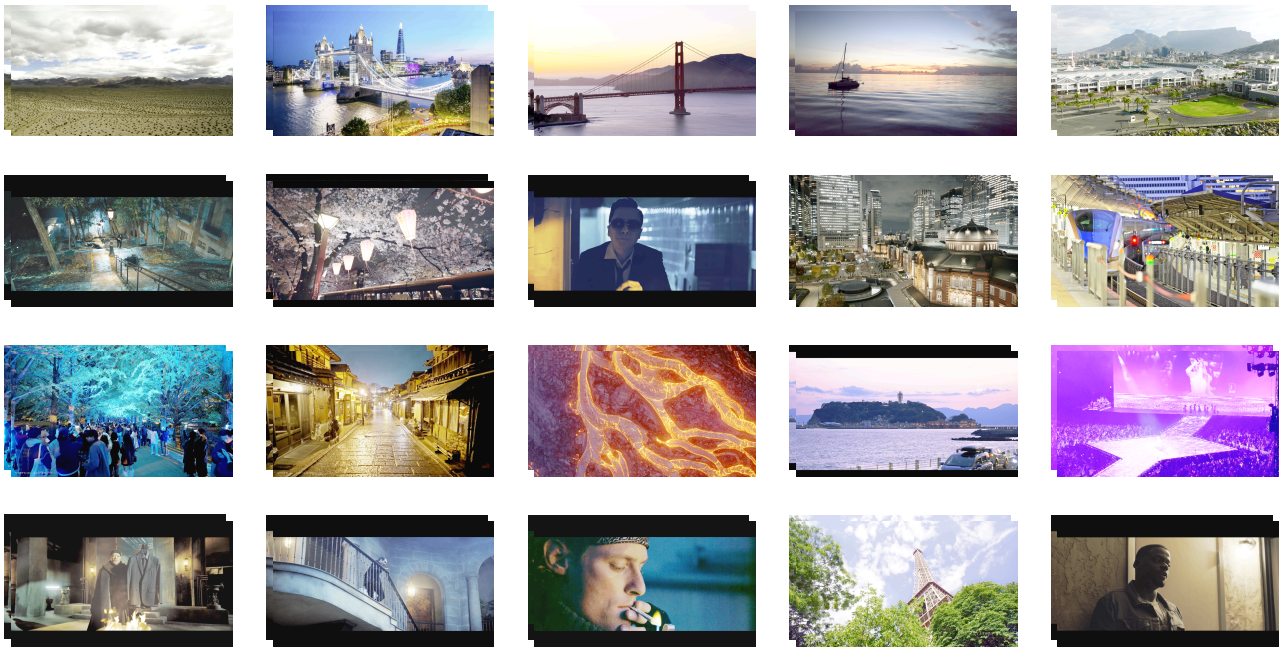

Figure 7. More video clips of HDRVD2K, which are all tone-mapped.

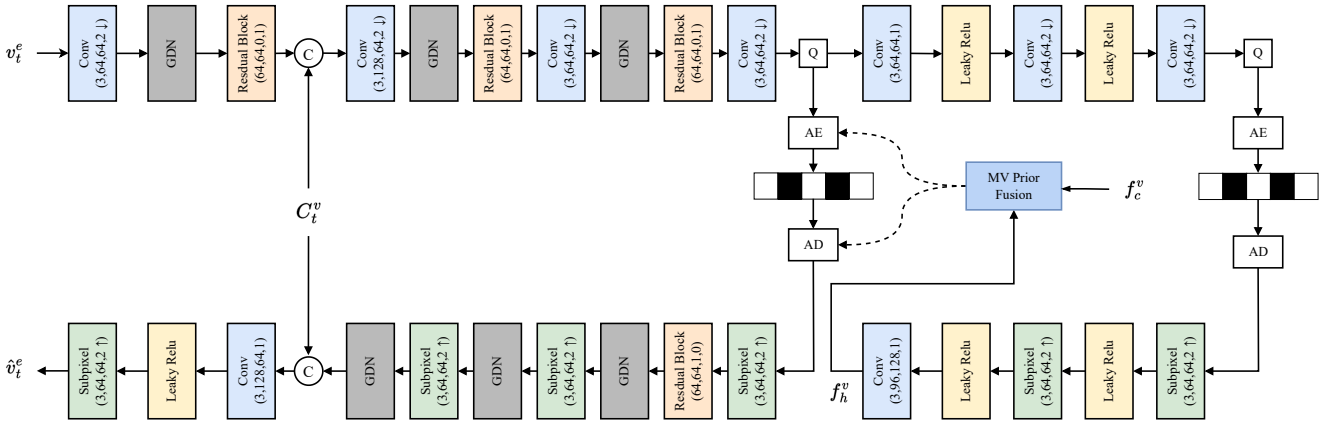

Figure 8. Architecture of contextual MV encoder-decoder. "Q" indicates the quantization. "AE" and "AE" indicate the arithmetic encoding and the arithmetic decoding. © indicates channel dimension concatenation.

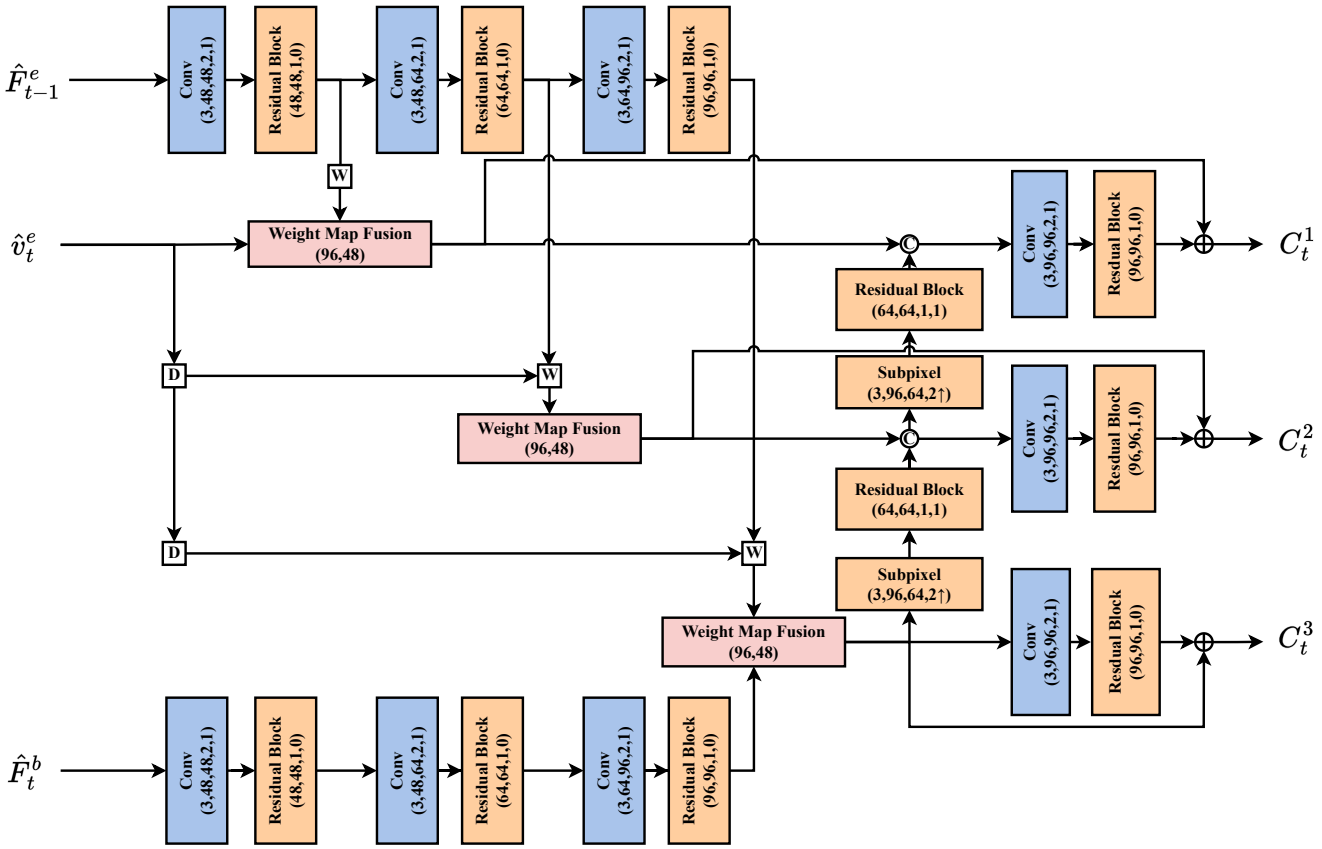

Figure 9. Architecture of HTLCM. The numbers in a "Weight Map Fusion" refer to the number of input channels and the number of output channels. "D" indicates bilinear downsampling. "W" indicates warping.

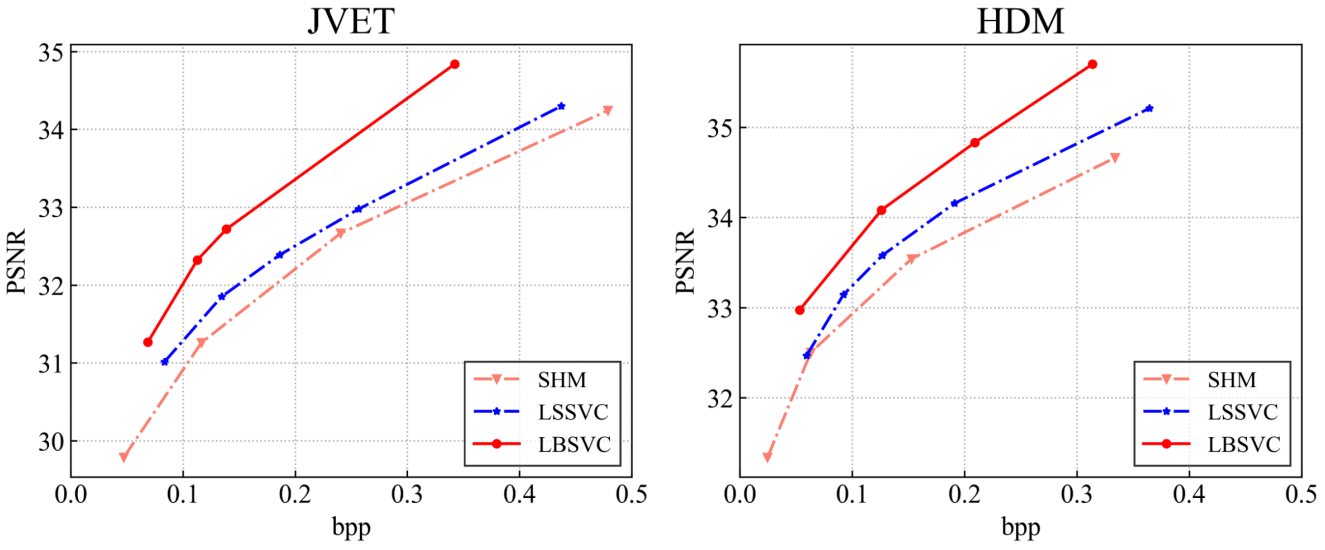

Figure 10. BL Rate-Distortion performance comparison on JVET and HDM dataset.