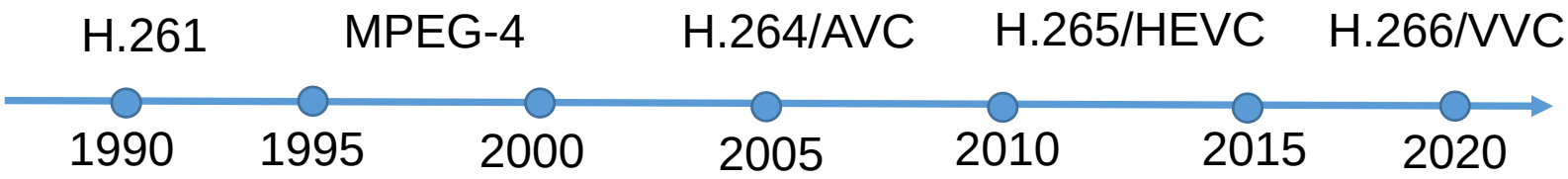
An Introduction of Deep Learning-Based Video Coding

Jianping Lin

2020/02/11

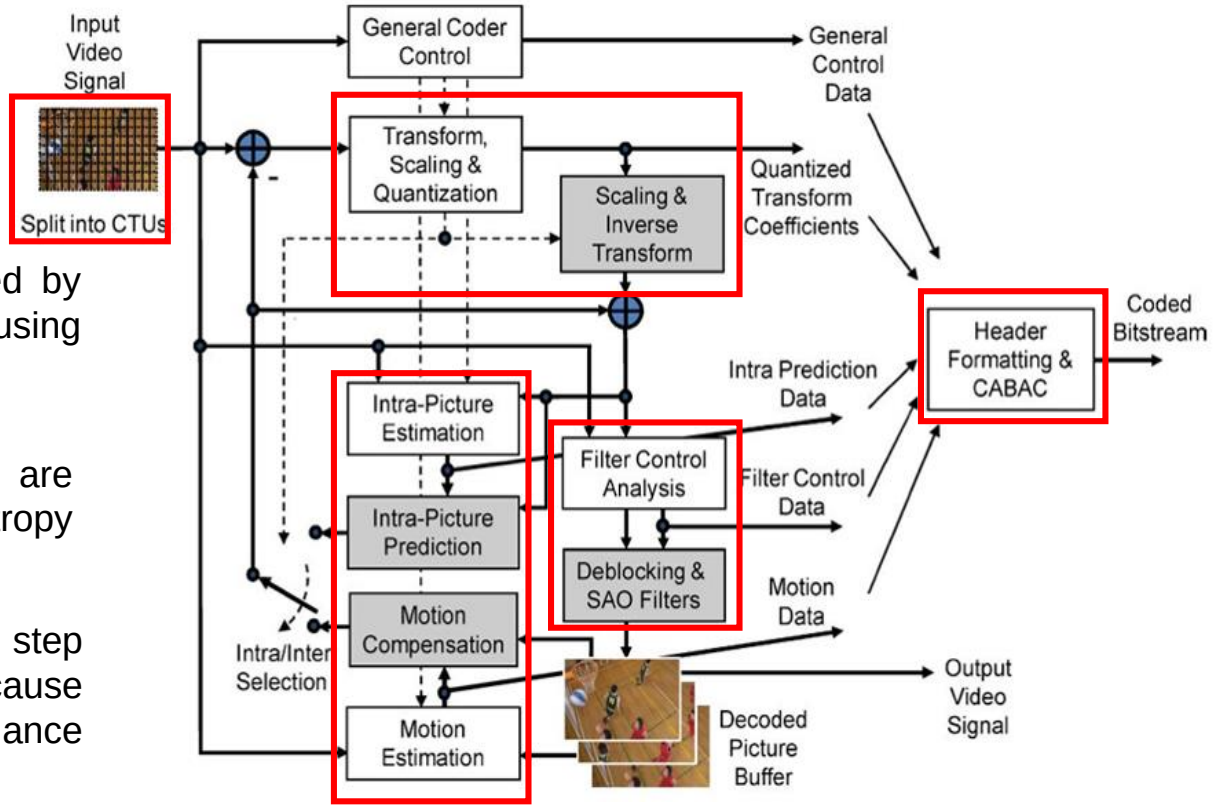
Traditional Video Coding Framework

- **Video coding standards**



- **All the standards in use or in the way coming follow the same traditional framework which combines predictive coding and transform coding**

- First, the current picture is divided into blocks, the largest block is called CTU.
- Then, the current block is predicted by the previously compressed blocks using intra/inter prediction.
- After that, the prediction residues are transformed and quantized and entropy coded to achieve the final bits.
- In addition, since the quantization step loses information and may cause artifacts, filtering is proposed to enhance the reconstructed pictures.

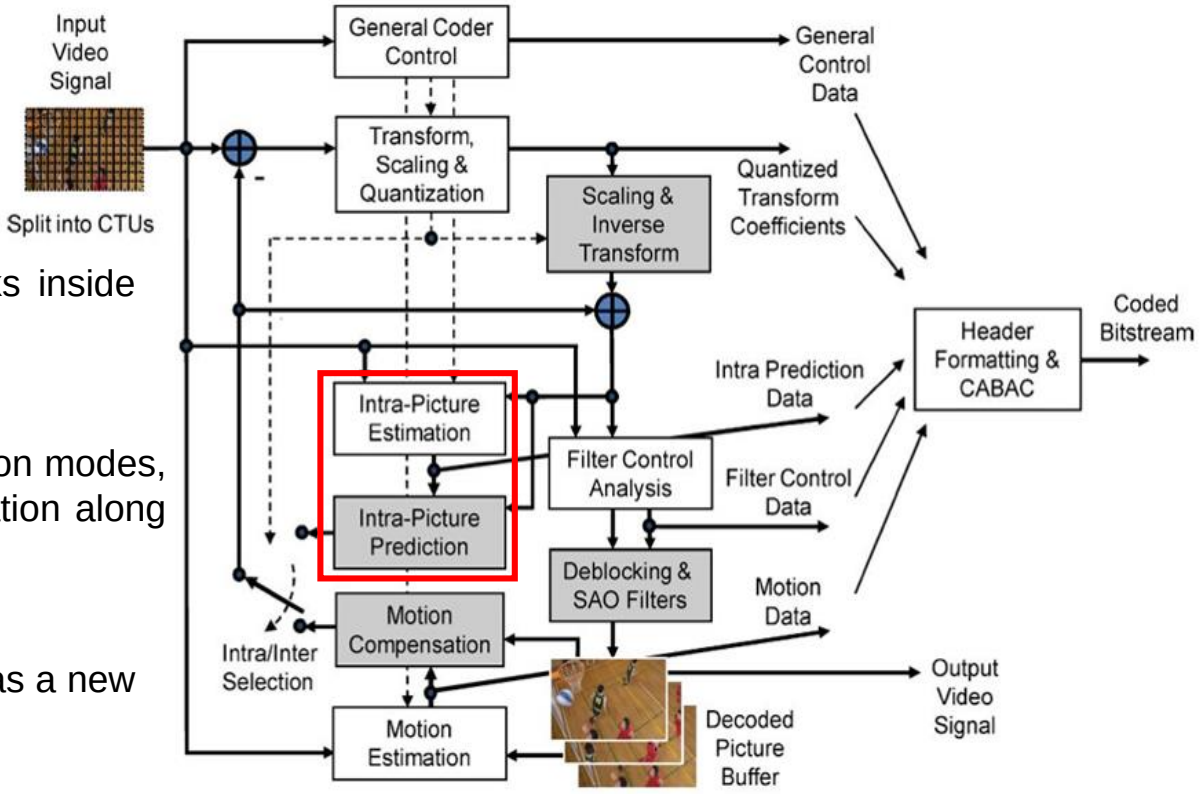


Deep Tools in Traditional Framework

- **Trained deep networks can act as almost all of the modules in traditional framework. The following are some of the deep tools.**

- **Intra Prediction in traditional framework**

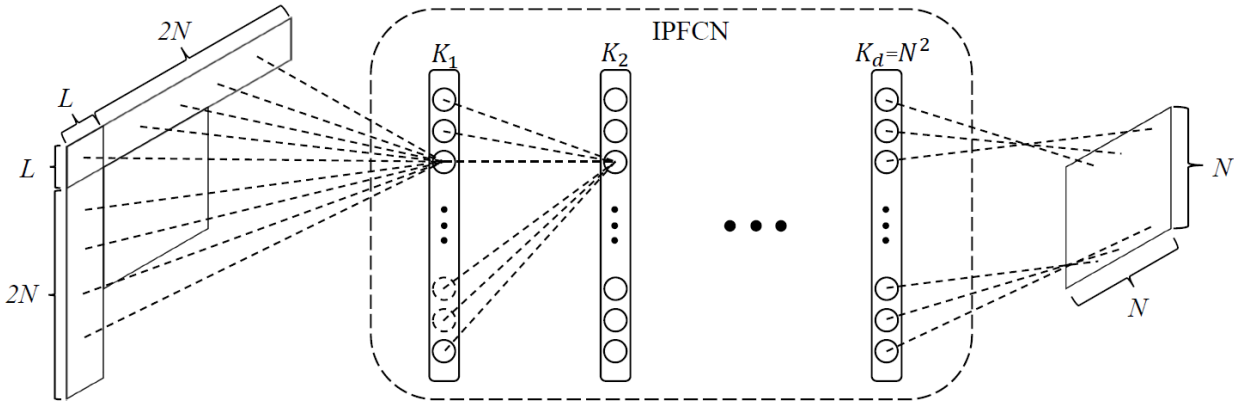
- It is used at the location indicated by the red box in the right figure.
- It is a tool to predict between blocks inside the same picture.
- There are several predefined prediction modes, such as DC prediction and extrapolation along different directions in HEVC
- Trained deep networks can be used as a new tool or together with existing modes.



Deep Tools in Traditional Framework

Deep tools for Intra Prediction

- Li et al. [1] propose a fully connected network for intra prediction that is shown in the right Figure.



- For the current N x N block, this network uses L rows above and L columns to the left, in total 4NL+L^2 compressed pixels as input to predict the current block.
- They use the HM (known as the test model of HEVC) to compress the raw images at different quantization parameters (QPs).
- When training the network, they split the training data into two groups by considering the HEVC prediction modes, and to train two models respectively.
- They integrate the trained networks as new prediction modes along with the HEVC modes, and achieve around 3% BD-rate reduction than HEVC.

[1] J. Li, B. Li, J. Xu, R. Xiong, and W. Gao, “Fully connected network-based intra prediction for image coding,” IEEE Transactions on Image Processing, vol. 27, no. 7, pp. 3236–3247, 2018.

Deep Tools in Traditional Framework

More deep tools for Intra Prediction

- J. Pfaff, P. Helle, D. Maniry, S. Kaltenstadler, W. Samek, H. Schwarz, D. Marpe, and T. Wiegand, “Neural network based intra prediction for video coding,” in Applications of Digital Image Processing XLI, vol. 10752. International Society for Optics and Photonics, 2018, p. 1075213.

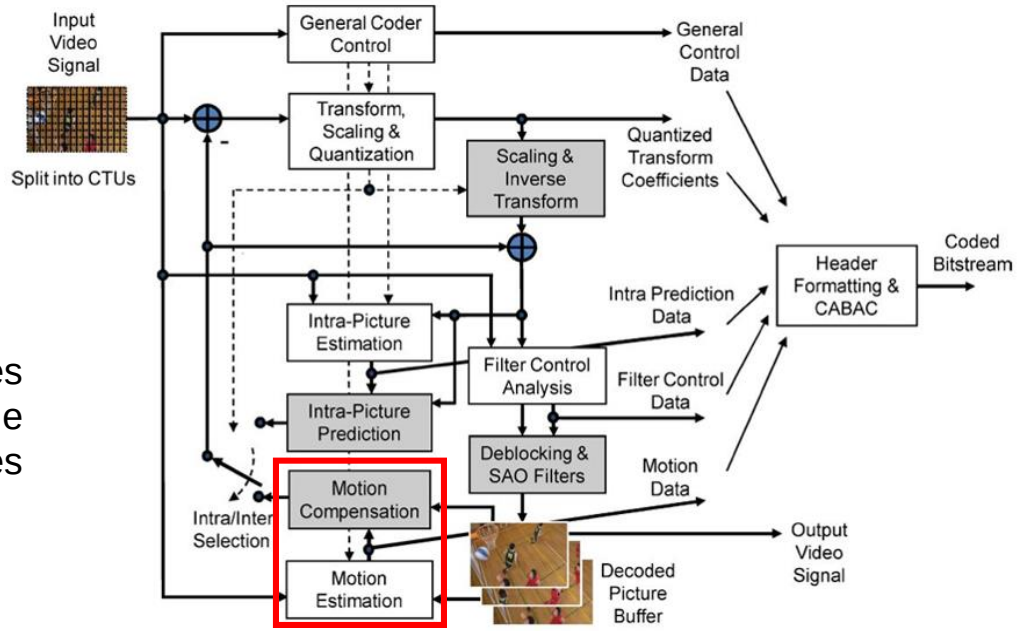
- Y. Hu, W. Yang, M. Li, and J. Liu, “Progressive spatial recurrent neural network for intra prediction,” arXiv preprint arXiv:1807.02232, 2018.

- W. Cui, T. Zhang, S. Zhang, F. Jiang, W. Zuo, Z. Wan, and D. Zhao, “Convolutional neural networks based intra prediction for HEVC,” in DCC. IEEE, 2017, p. 436.

Deep tools in Traditional Framework

☐ Inter Prediction in traditional framework

- It is used at the location indicated by the red box in the right figure.



- It is a tool to predict between video frames so as to remove the redundancy along the temporal dimension, and it largely decides the compression efficiency.

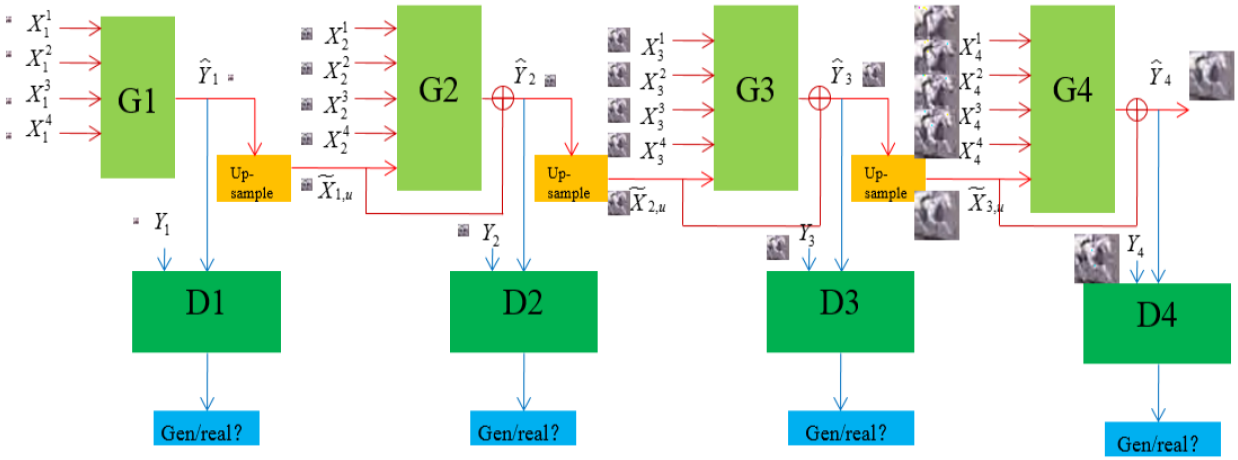
- It is mostly fulfilled by block-level motion estimation (ME) and motion compensation (MC).

Specifically, given a reference frame and a block to be coded, ME is to find the location in the reference frame where the content is the most similar to that inside the to-be-coded block, and MC is to retrieve the content at the found location so as to predict the block.

- There have been many deep tools developed to improve block-level ME and MC in traditional framework.

Deep tools in traditional framework

Deep tools for Inter Prediction



- Inspired by the multiple reference frames, Lin et al. [2] propose a new inter prediction mechanism by extrapolating the next one frame.
- As shown in this figure, they adopt a Laplacian pyramid of GANs to extrapolate a frame from the previously compressed four frames.
- This extrapolated frame serves as another reference frame in HM, and achieves around 2% BD-rate reduction than HEVC.

[2] J. Lin, D. Liu, H. Li, and F. Wu, “Generative adversarial network-based frame extrapolation for video coding,” in VCIP, 2018, pp. 1–4.

Deep Tools in Traditional Framework

More deep tools for Inter Prediction

- 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 Transactions on Circuits and Systems for Video Technology, DOI: 10.1109/TCSVT.2018.2876399, 2018.

- N. Yan, D. Liu, H. Li, and F. Wu, “A convolutional neural network approach for half-pel interpolation in video coding,” in ISCAS. IEEE, 2017, pp. 1–4.

- H. Zhang, L. Song, Z. Luo, and X. Yang, “Learning a convolutional neural network for fractional interpolation in HEVC inter coding,” in VCIP. IEEE, 2017, pp. 1–4.

- N. Yan, D. Liu, B. Li, H. Li, T. Xu, and F. Wu, “Convolutional neural network-based invertible half-pixel interpolation filter for video coding,” in ICIP, 2018, pp. 201–205.

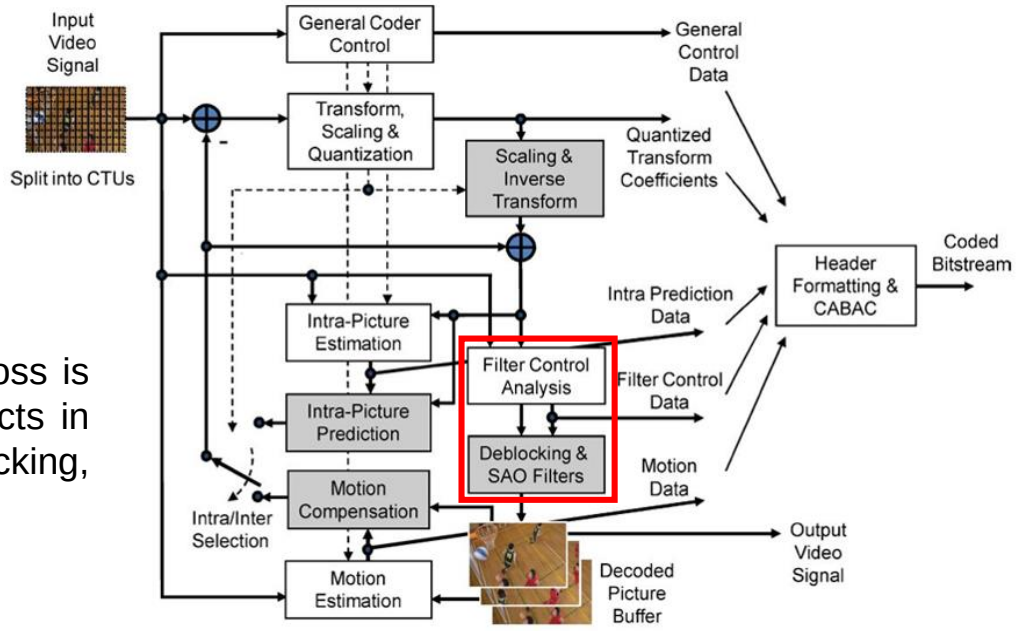
- S. Huo, D. Liu, F. Wu, and H. Li, “Convolutional neural network-based motion compensation refinement for video coding,” in ISCAS, 2018, pp. 1–4.

- Y. Wang, X. Fan, C. Jia, D. Zhao, and W. Gao, “Neural network based inter prediction for HEVC,” in ICME. IEEE, 2018, pp. 1–6.

Deep tools in Traditional Framework

Filtering in traditional framework

- Most of the widely used video coding schemes are lossy coding ones.



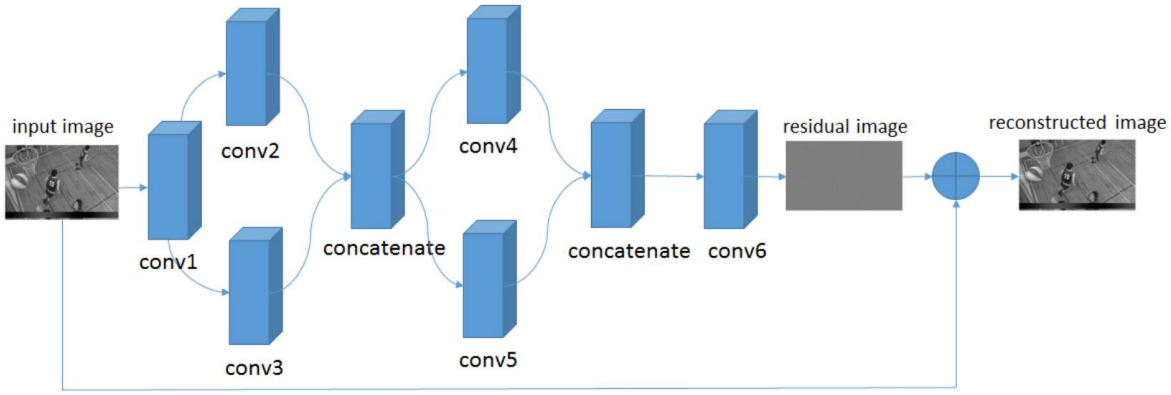
- When the quantization step is large, the loss is large too, which may lead to visible artifacts in the reconstructed pictures, such as blocking, blurring, ringing, color shift, and flickering.

- Filtering is the tool to reduce these artifacts. In HEVC, two in-loop filters are presented in this figure, namely deblocking filter (DF) and sample adaptive offset (SAO).

- Many deep tools have been developed to replace or be used together with the existing filters.

Deep tools in traditional framework

Deep tools for filtering



Layer	Layer 1	Layer 2		Layer 3		Layer 4
Conv. module	conv1	conv2	conv3	conv4	conv5	conv6
Filter size	5×5	5×5	3×3	3×3	1×1	3×3
# filters	64	16	32	16	32	1
# parameters	1600	25600	18432	6912	1536	432
Total parameters	54512					

- Dai et al. [3] propose a 4-layer CNN for filtering of reconstructed frames in HEVC, where the CNN has variable filter size and residue connection, and named VRCNN.
- They build the training set by using the HM to compress raw images at different QPs, and use the following loss function to train different models for different QPs.

$$L(\Theta) = \frac{1}{N}\sum_{n=1}^{N} \|F(\boldsymbol{Y}_n|\Theta) - \boldsymbol{X}_n\|^2$$

- When integrating the trained VRCNN into HEVC, the deblocking and SAO are turned off. It achieves 4.6% BD-rate reduction than HEVC.

[3] Y. Dai, D. Liu, and F. Wu, “A convolutional neural network approach for post-processing in HEVC intra coding,” in MMM. Springer, 2017, pp. 28–39.

Deep Tools in Traditional Framework

More deep tools for filtering

- T. Wang, M. Chen, and H. Chao, “A novel deep learning-based method of improving coding efficiency from the decoder-end for HEVC,” in DCC. IEEE, 2017, pp. 410–419.
- R. Yang, M. Xu, T. Liu, Z. Wang, and Z. Guan, “Enhancing quality for HEVC compressed videos,” IEEE Transactions on Circuits and Systems for Video Technology, DOI: 10.1109/TCSVT.2018.2867568, 2018.
- Z. Jin, P. An, C. Yang, and L. Shen, “Quality enhancement for intra frame coding via CNNs: An adversarial approach,” in ICASSP. IEEE, 2018, pp. 1368–1372.
- R. Yang, M. Xu, Z. Wang, and T. Li, “Multi-frame quality enhancement for compressed video,” in CVPR, 2018, pp. 6664–6673.
- W.-S. Park and M. Kim, “CNN-based in-loop filtering for coding efficiency improvement,” in IEEE Image, Video, and Multidimensional Signal Processing Workshop. IEEE, 2016, pp. 1–5.
- X. Meng, C. Chen, S. Zhu, and B. Zeng, “A new HEVC in-loop filter based on multi-channel long-short-term dependency residual networks,” in DCC. IEEE, 2018, pp. 187–196.
- C. Jia, S. Wang, X. Zhang, S. Wang, J. Liu, S. Pu, and S. Ma, “Content-aware convolutional neural network for in-loop filtering in high efficiency video coding,” IEEE Transactions on Image Processing, DOI: 10.1109/TIP.2019.2896489, 2019.

Deep Coding Schemes

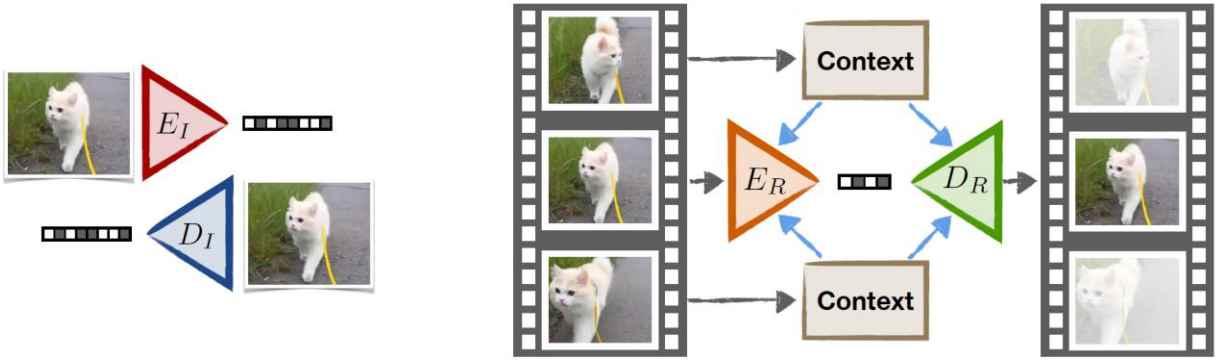
- **The previous deep tools are acting as the new modules in traditional framework. The following are some deep schemes that can be used to compress videos independently.**

  - Deep schemes for random-access scenarios like playback
    - Random Access means every concession of several frames can be decoded independently, but the latency is longer. These schemes are usually based on frame interpolation which uses the previous and the subsequent frames as references to compress the current frame.
    - Wu_ECCV2018 and Djelouah_ICCV2019 are two typical works for this scenarios.

  - Deep schemes for low-latency scenarios like live transmission
    - These schemes restrict the networks to use only temporally previous frames as references, and thus the latency can be very low.
    - Lu_CVPR2019 and Rippel_ICCV2019 are two typical works for this scenarios.

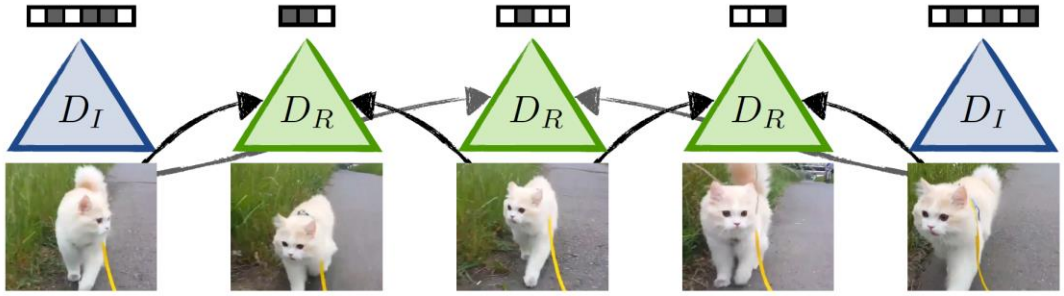
Deep Coding Schemes

☐ Deep schemes for random-access scenarios: Wu_ECCV2018 [4]



(a) I-frame model. (b) Final interpolation model.

- Wu et al. [4] proposed a model composed of an image compression model that compresses the key frames (Figure a), and a conditional interpolation model that compresses the remaining frames (Figure b).



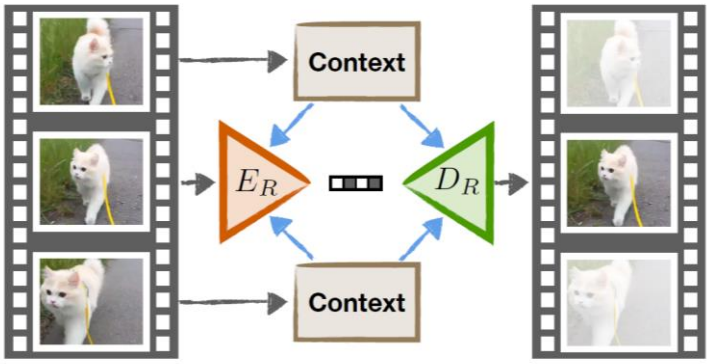
- They compress all frames in a hierarchical order, as shown in this figure for 5 frames. The first and last frames are compressed by I-frame model, and then the middle frame are compressed by one interpolation model. Finally, the remaining two frames are compressed by another interpolation model.

[4] C.-Y. Wu, N. Singhal, and P. Kr¨ahenb¨uhl, “Video compression through image interpolation,” in ECCV, 2018, pp. 416–431.

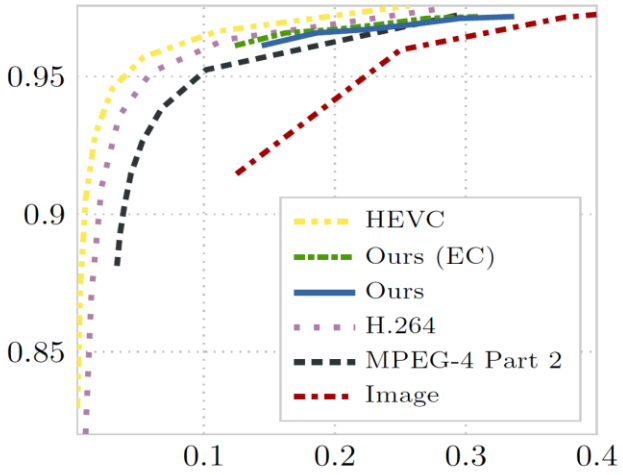
Deep Coding Schemes

☐ Deep schemes for random-access scenarios： Wu_ECCV2018 [5]

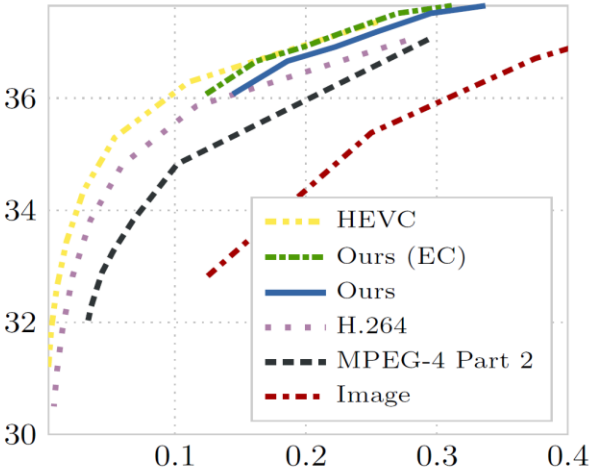
- However, for the interpolation model, the motion information used to warp the before and after reconstructed frames are extracted by traditional block-level ME technique, such as the ME of H.264.



(b) Final interpolation model.



(a) MS-SSIM

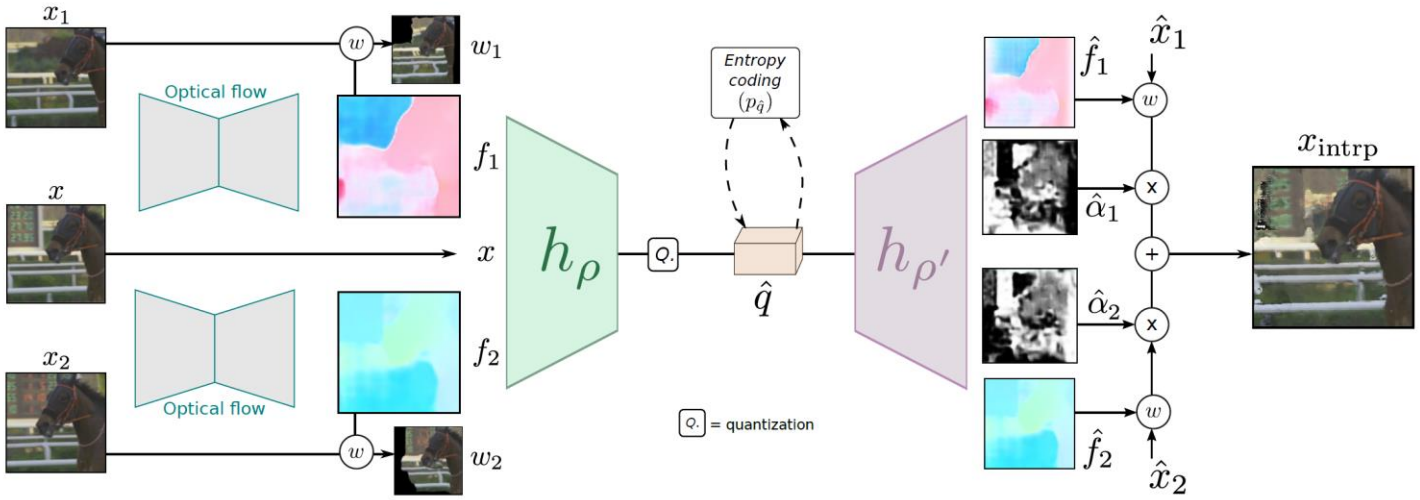


(b) PSNR (dB)

- Since the motion information is extracted by traditional technique and not from joint learning, the coding performance of this method is just on par with H.264.

Deep Coding Schemes

☐ Deep schemes for random-access scenarios： Djelouah_ICCV2019 [5]



- In 2019, Djelouah et al. [5] also proposed a method for interpolation-based video compression, where the interpolation model combines motion information compression and image synthesis.

- As shown in this figure, the motion information is extracted by an optical flow network. Then the original frame x, the optical flow fields (f_1, f_2) and the warped frames (w_1, w_2) are fed into the encoder. The decoder directly synthesize the displacement maps $(\hat{f}_1, \hat{f}_2)$, and the blending coefficients $(\hat{\alpha}_1, \hat{\alpha}_2)$ to compute the interpolated frame x_{intrp} via:

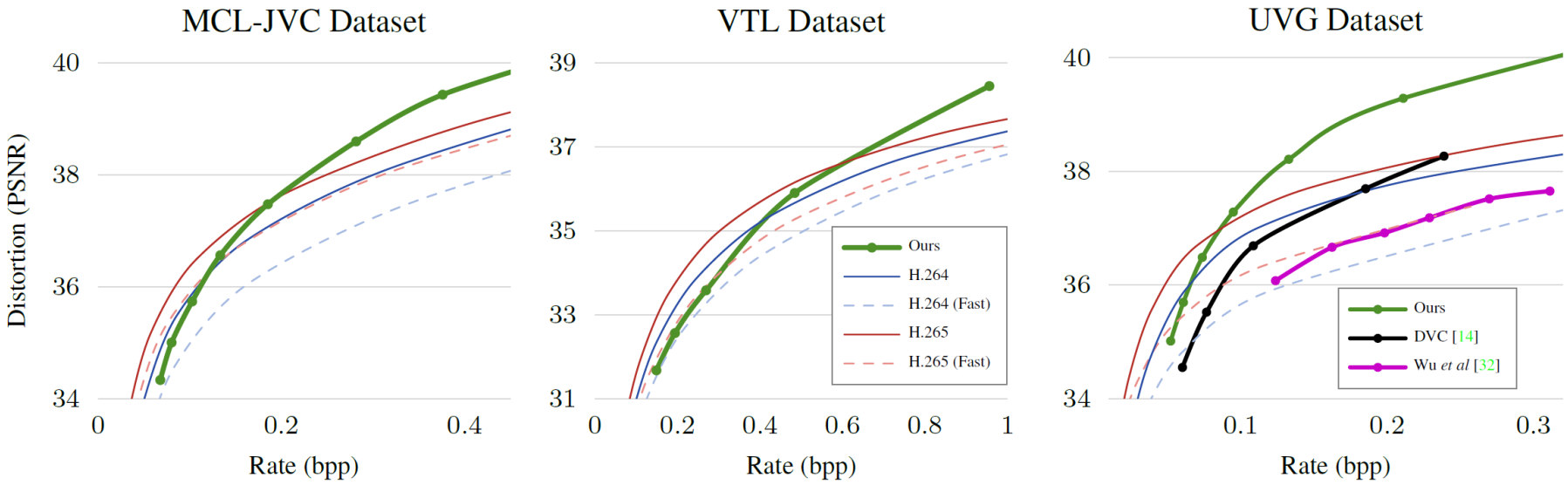
$$x_{\text{intrp}} = \sum_{i=1}^{k} \hat{\alpha}_i w(x_i, \hat{f}_i) \quad \text{with} \quad \sum_{i=1}^{k} \hat{\alpha}_i = 1$$

- Finally, the residual between x and x_{intrp} is compressed by another auto-encoder.

[5]Abdelaziz Djelouah, Joaquim Campos, Simone Schaub-Meyer, and Christopher Schroers. Neural inter-frame compression for video coding. In ICCV, October 2019.

Deep Coding Schemes

☐ Deep schemes for random-access scenarios： Djelouah_ICCV2019 [5]



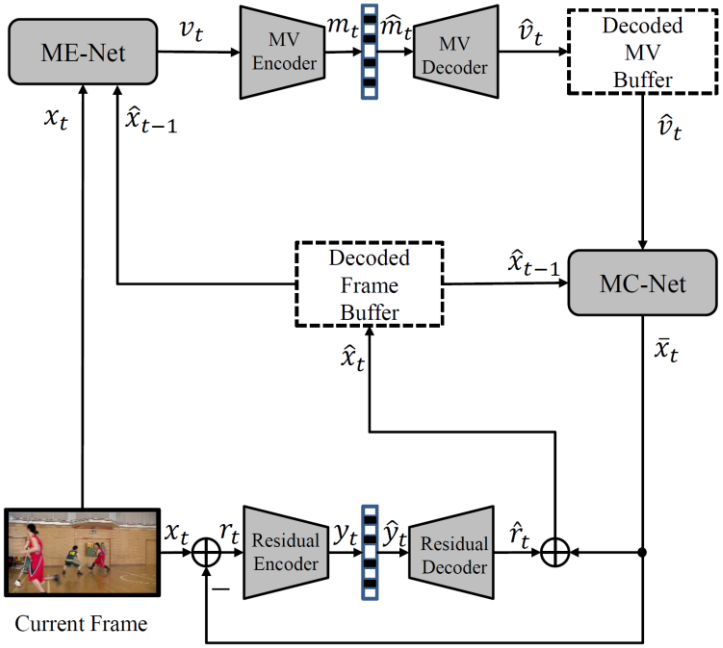
- After a carefully training, this method has performed better than H.265 in PSNR at high bit-rate range, which is the best compression performance among all learning-based methods for random-access mode.

[5]Abdelaziz Djelouah, Joaquim Campos, Simone Schaub-Meyer, and Christopher Schroers. Neural inter-frame compression for video coding. In ICCV, October 2019.

Deep Coding Schemes

☐ Deep schemes for low-latency scenarios: Lu_CVPR2019 [6]

- In CVPR2019, Lu et al. [6] proposed a method for low-latency video compression, where the network is restricted to use the temporally previous one frame as references.



- Specifically, they feed the previous reference frame $\hat{x}_{t-1}$ and the current original frame x_t into an optical flow network to extract optical flow field as motion information v_t . And then v_t is compressed by an auto-encoder. After that, the motion compensation network use the compressed $\hat{v}_t$ and $\hat{x}_{t-1}$ to obtain the prediction signal $\bar{x}_t$. Finally, the residual r_t between x_t and $\bar{x}_t$ is compressed by another auto-encoder.

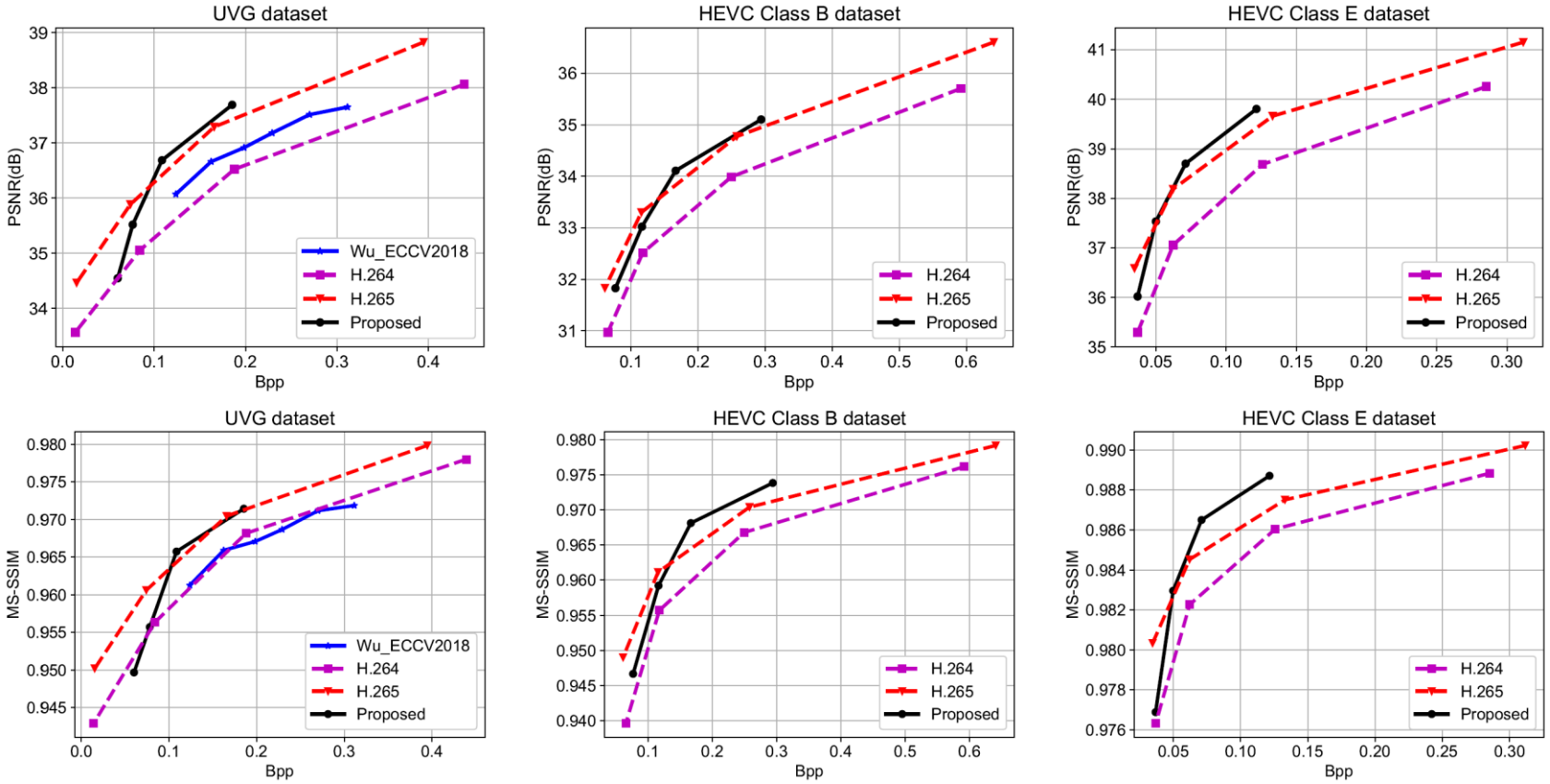
- All modules are jointly optimized by a single rate-distortion loss function:

$$\lambda D + R = \lambda d(x_t, \hat{x}_t) + (H(\hat{m}_t) + H(\hat{y}_t))$$

[6] Guo Lu, Wanli Ouyang, Dong Xu, Xiaoyun Zhang, Chunlei Cai, and Zhiyong Gao. Dvc: An end-to-end deep video compression framework. In CVPR, June 2019.

Deep Coding Schemes

☐ Deep schemes for low-latency scenarios： Lu_CVPR2019 [6]



- Their method has outperformed H.264 in PSNR and MS-SSIM and achieved similar or better compression performance when compared with H.265 in terms of MS-SSIM. Up to now, it is the best compression performance in PSNR among all learning-based methods for low-latency mode.

[6] Guo Lu, Wanli Ouyang, Dong Xu, Xiaoyun Zhang, Chunlei Cai, and Zhiyong Gao. Dvc: An end-to-end deep video compression framework. In CVPR, June 2019.

M-LVC: Multiple Frames Prediction for Learned Video Compression

- **The framework of our end-to-end learned video compression**

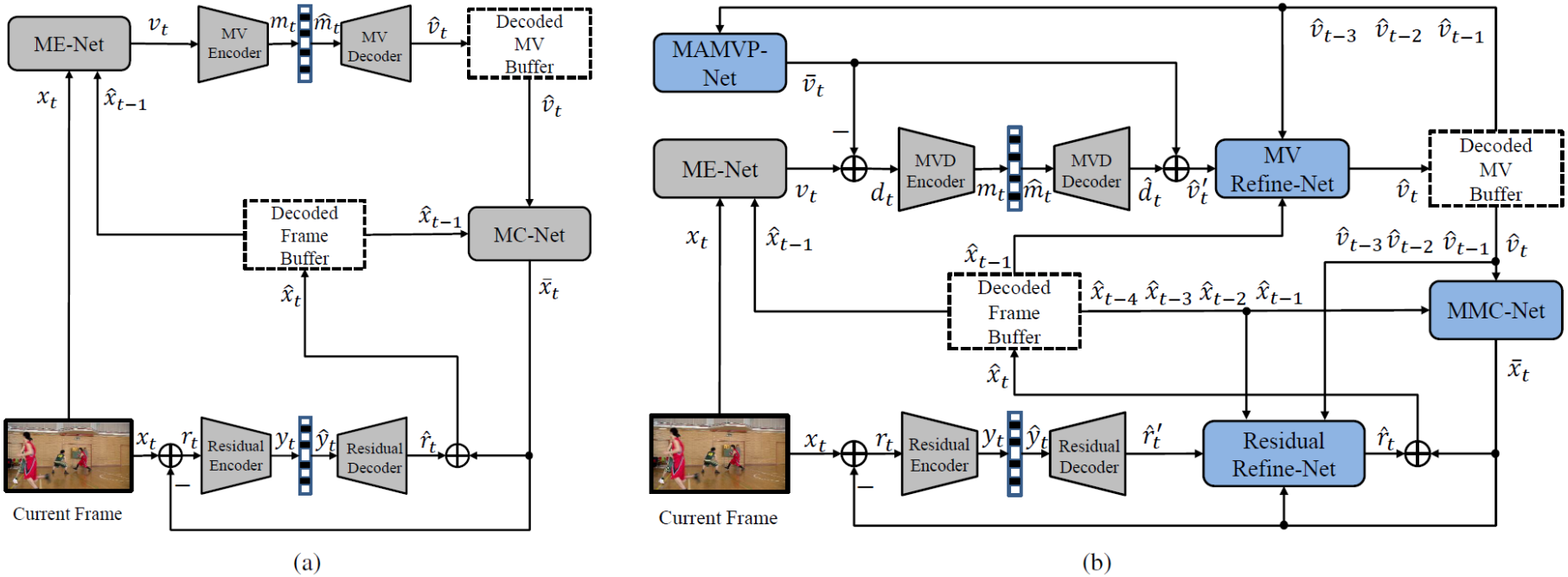


Figure 1. (a) The framework of DVC [14]. (b) Our framework. Compared to DVC, our framework has four new modules that are highlighted in blue. In addition, our Decoded Frame Buffer stores multiple previously decoded frames as references. Our Decoded MV Buffer also stores multiple decoded MV fields. Four reference frames are depicted in the figure, which is the default setting in this paper.

M-LVC: Multiple Frames Prediction for Learned Video Compression

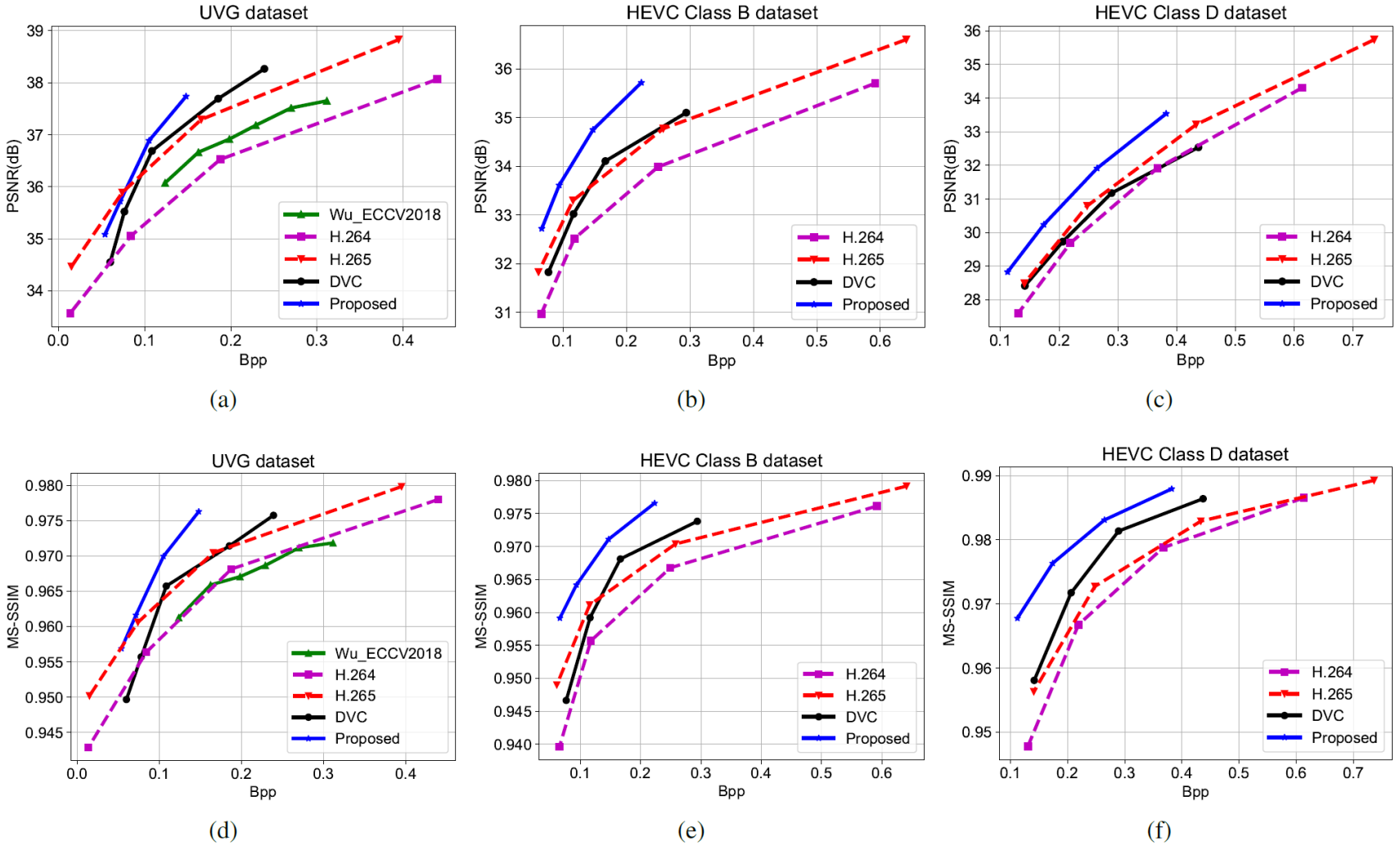


Figure 4. **Overall performance**. The compression results on three test sets using H.264 [28], H.265 [20], DVC [14], Wu's method [29] and the proposed method. We directly use the results reported in [14] and [29]. Wu [29] did not report on HEVC Class B and Class D. Top row: PSNR. Bottom row: MS-SSIM.

M-LVC: Multiple Frames Prediction for Learned Video Compression

Table 1. The average time per frame of encoding and decoding 320x256 videos using our different models.

Model	Our Baseline	Add MAMVP-Net	Add MVRefine-Net	Add MMC-Net	Proposed
Encoding Time	0.25s	0.31s	0.34s	0.35s	0.37s
Decoding Time	0.05s	0.11s	0.14s	0.15s	0.17s


Thank
You!