An Introduction of Learned Video Compression

Jianping Lin

2020/03/04

Learned Video Compression

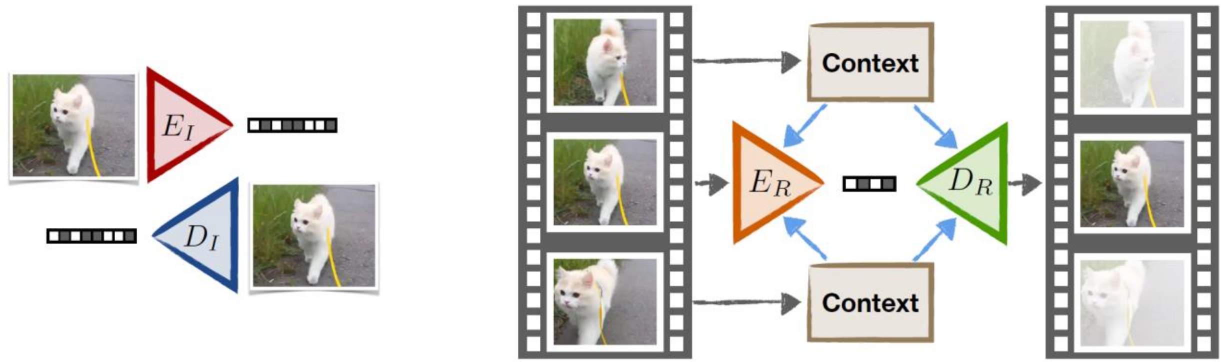
- **Learned video compression is the scheme built on neural networks, which can be used to compress videos individually.**

  - Deep schemes for random-access scenarios like playback
    - Random Access means each group of frames (GoP) can be decoded independently, but the latency is longer. These schemes are usually based on frame interpolation which uses the before and the after 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, Rippel_ICCV2019, and Liu_AAAI2020 are typical works for this scenarios.

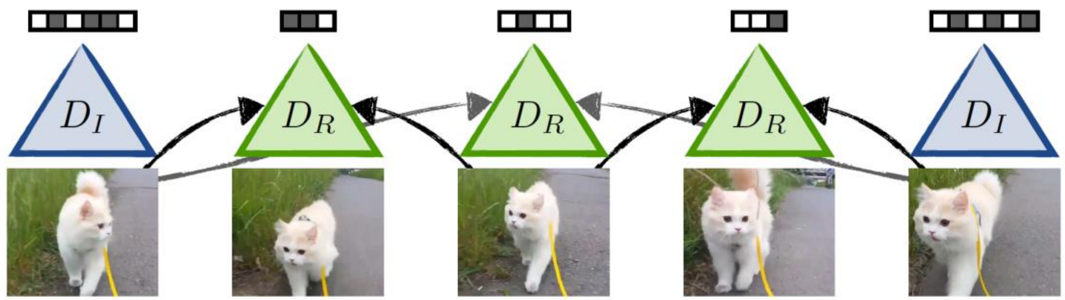
Learned Video Compression

☐ Deep schemes for random-access scenarios： Wu_ECCV2018 [1]



(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).



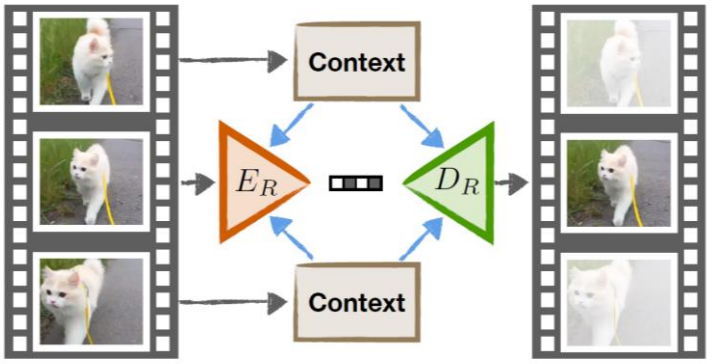
- They compress all frames in a hierarchical order, as shown in this figure for 5 frames. The first and last frames are compressed by I-frame model, and then the middle frame is compressed by one interpolation model. Finally, the remaining two frames are compressed by another interpolation model.

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

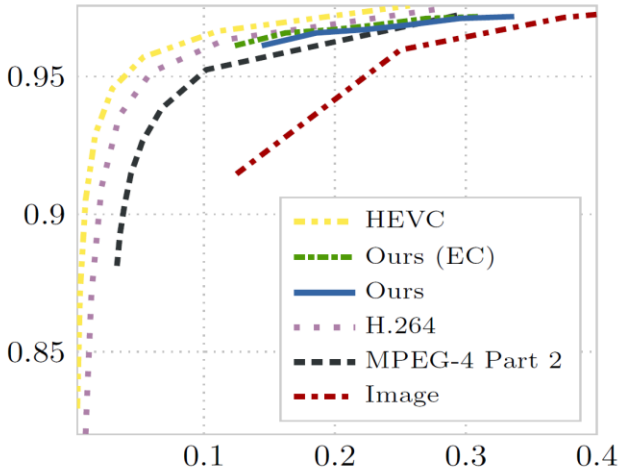
Learned Video Compression

☐ Deep schemes for random-access scenarios： Wu_ECCV2018 [1]

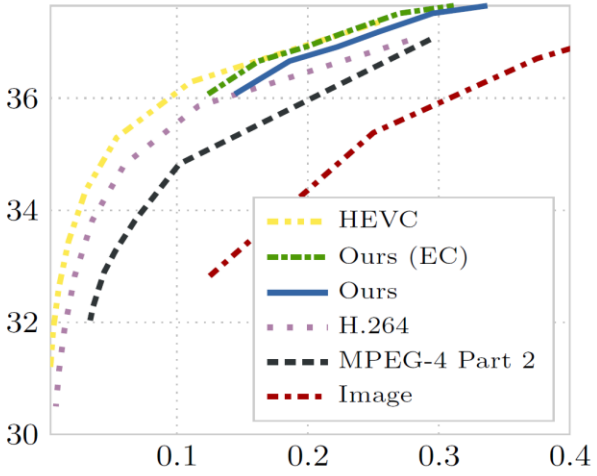
- However, the motion information used to warp the before and after reconstructed frames are extracted by traditional block-level ME technique, i.e. the ME of H.264.



(b) Final interpolation model.



(a) MS-SSIM

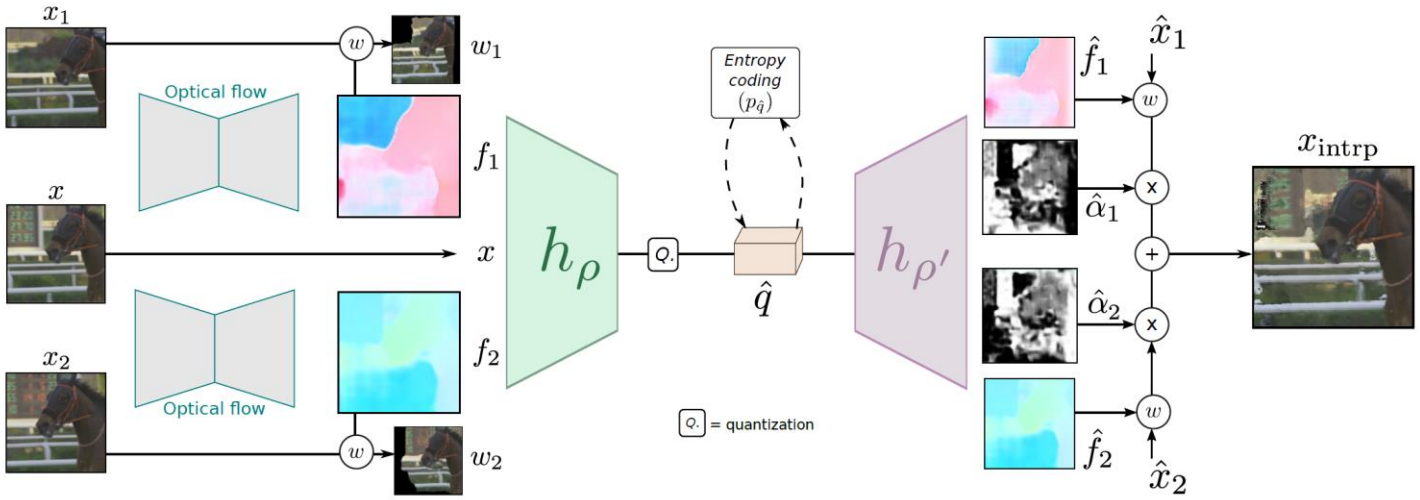


(b) PSNR (dB)

- 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.

Learned Video Compression

☐ Deep schemes for random-access scenarios: Djelouah_ICCV2019 [2]



- 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:

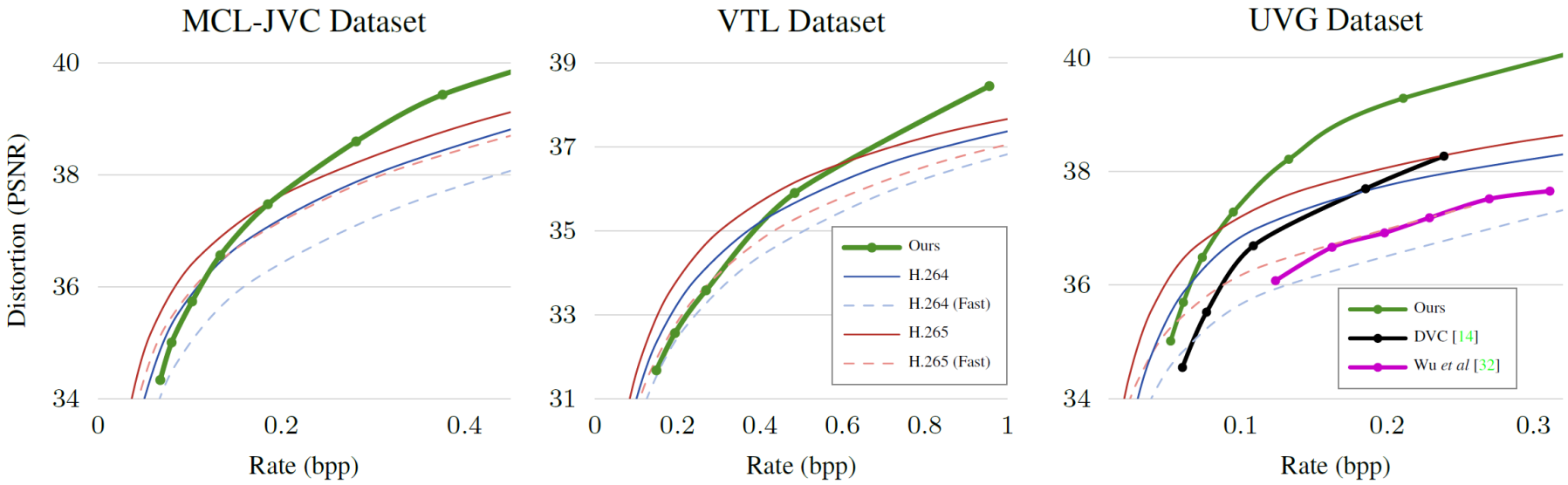
$$x_{\text{intrp}} = \sum_{i=1}^{k} \hat{\alpha}_i w(x_i, \hat{f}_i) \quad \text{with} \quad \sum_{i=1}^{k} \hat{\alpha}_i = 1$$

- Finally, the residual between x and x_{intrp} is compressed by another auto-encoder which uses both hyperprior and autoregressive model.

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

Learned Video Compression

◘ Deep schemes for random-access scenarios： Djelouah_ICCV2019 [2]



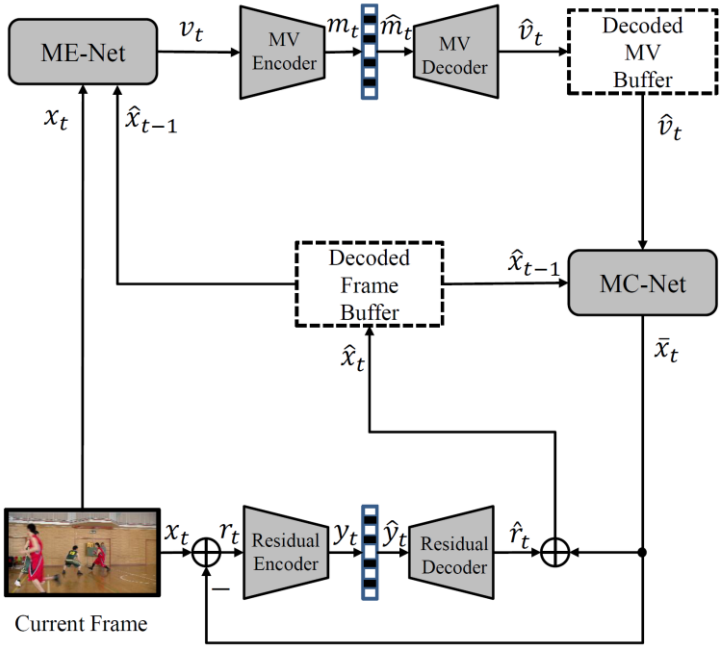
- After a carefully training, this method has performed better than H.265 in PSNR at high bit-rate range, which is the best compression performance among all learning-based methods for random-access mode.

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

Learned Video Compression

☐ Deep schemes for low-latency scenarios: Lu_CVPR2019 [3]

- 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.



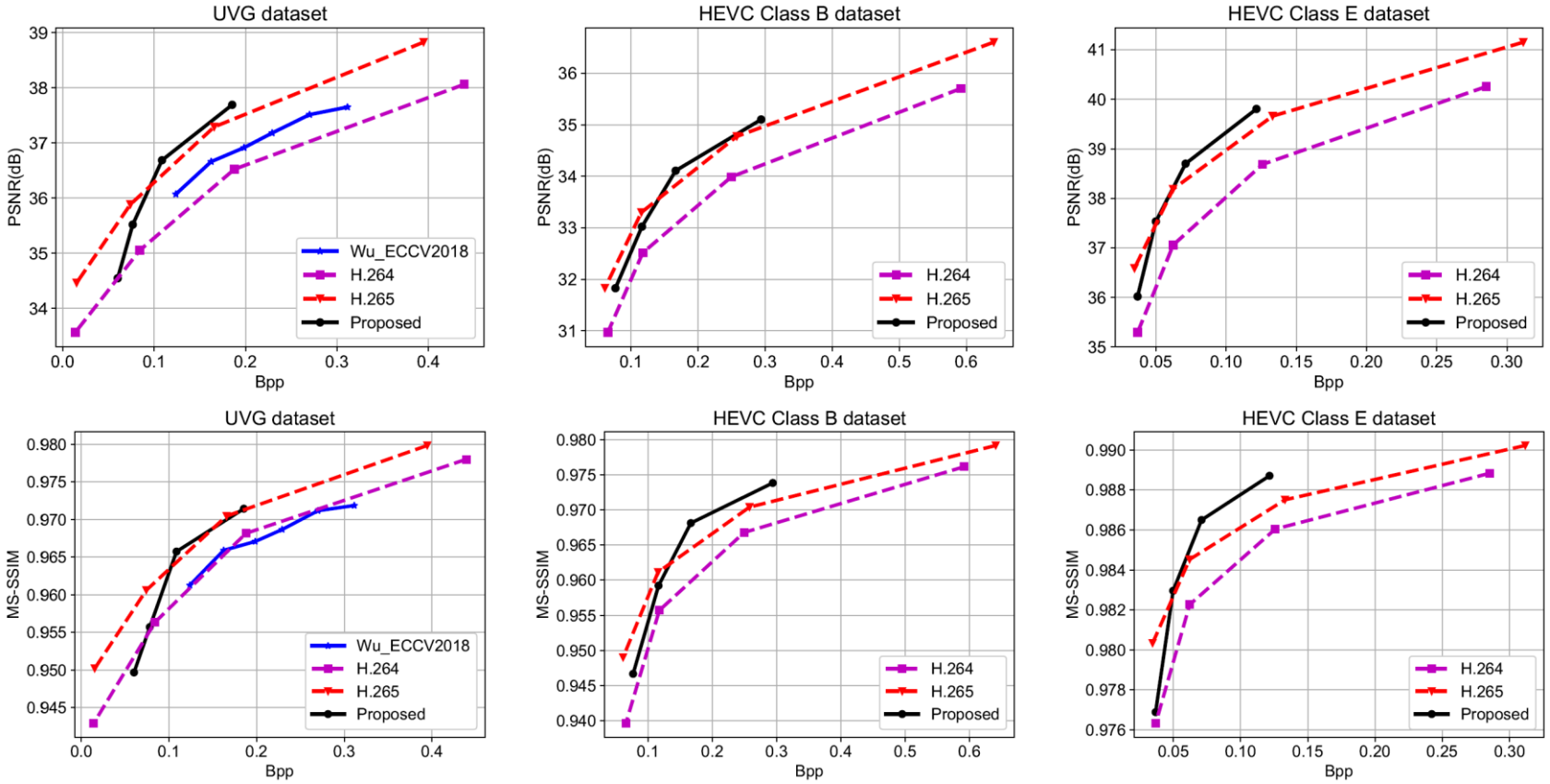
- Specifically, they feed the previous reference frame $\hat{x}_{t-1}$ and the current original frame x_t into an optical flow network to extract optical flow field as motion information v_t . And then v_t is compressed by an fully-factorized auto-encoder. After that, the motion compensation network 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 hyperprior-based auto-encoder.
- All modules are jointly optimized by a single rate-distortion loss function:

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

[3] 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.

Learned Video Compression

☐ Deep schemes for low-latency scenarios： Lu_CVPR2019 [3]

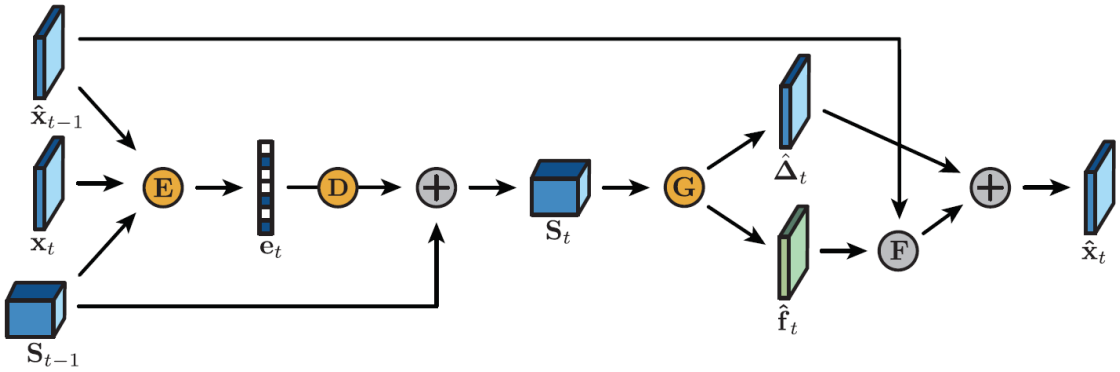


➢ Their method has outperformed H.264 in PSNR and MS-SSIM and achieved similar or better compression performance when compared with H.265 in terms of MS-SSIM.

[3] 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.

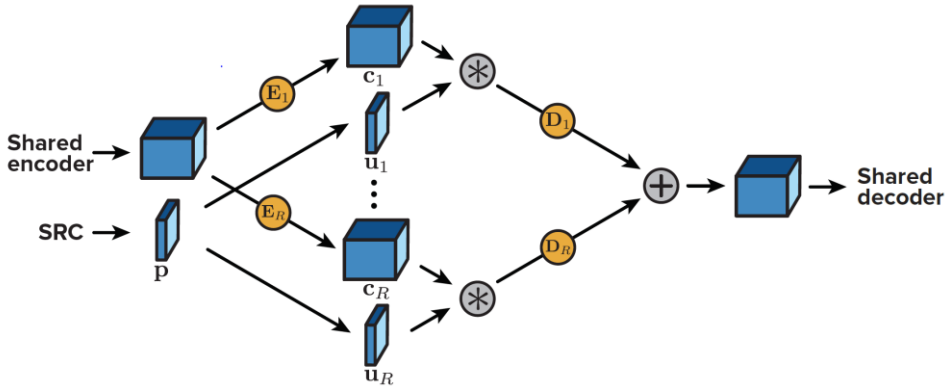
Learned Video Compression

☐ Deep schemes for low-latency scenarios： Rippel_ICCV2019 [4]



- They proposed a method for low-latency video compression, where the network maintains a latent state to accumulates temporal information through recursive updates like a RNN.
- The auto-encoder is used to compress the motion information and the residual simultaneously.

- They also proposed a spatial rate control algorithm to independently assign arbitrary bitrates at different spatial locations. It can assign more bits to the areas that are harder to reconstruct.

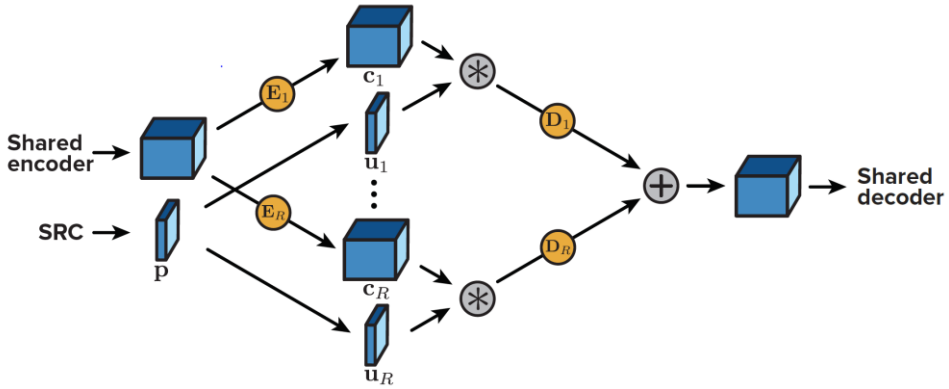


The architecture of the spatial multiplexer for rate control

[4] 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.

Learned Video Compression

Deep schemes for low-latency scenarios: Rippel_ICCV2019 [4]



- First, they map the spatial map of integer rates $p \in \{1,2,\ldots,R\}^{YxX}$ into R binary masks $u_r \in \{0,1\}^{C_r x Y_r x X_r}$, via:

$$\mathrm{u}_{r,cyx} = \mathbb{I}_{\mathrm{p}_{yx}=r}\,, \qquad r = 1,\ldots,R$$

- Each map u_r masks codelayer c_r during entropy coding. The final bitstream then corresponds to encodings of all the active values in each codelayer, as well as the rate mask itself.

- During training, they simply sample p uniformly for each frame.

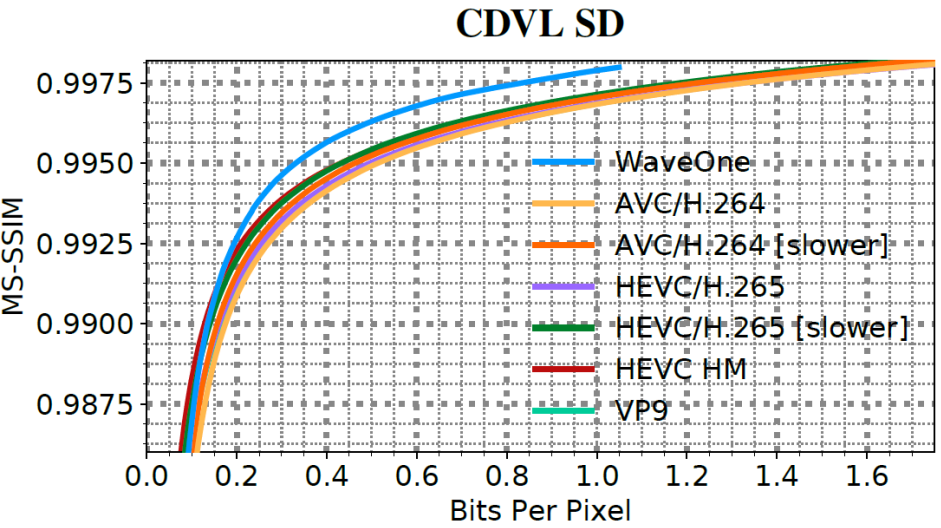
- During deployment, they estimate the slope $\frac{\mathscr{L}_{r+1,yx}-\mathscr{L}_{r,yx}}{\mathrm{BPP}_{r+1,yx}-\mathrm{BPP}_{r,yx}}$ of the local R-D curve for each location (y,x) and rate r. They then choose the rate map p such that at each location p_{yx} is the largest rate such that the slope is at least the predefined threshold λ.

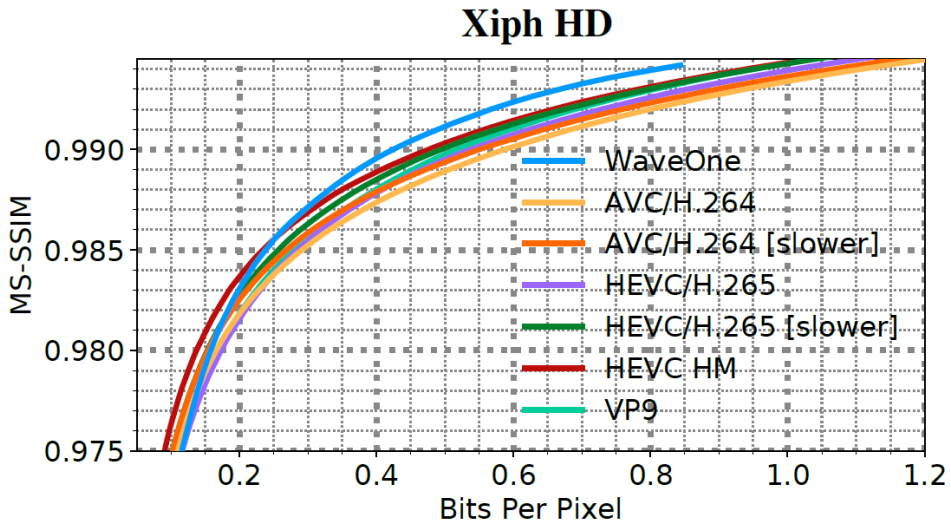
- Their experimental results showed that the spatial rate controller achieved 10-20% better compression.

[4] Oren Rippel, Sanjay Nair, Carissa Lew, Steve Branson, Alexander G. Anderson, and Lubomir Bourdev. Learned video compression. In ICCV, October 2019. 1, 2, 3, 6

Learned Video Compression

☐ Deep schemes for low-latency scenarios： Rippel_ICCV2019 [4]





- Their method has outperformed HEVC HM in terms of MS-SSIM at high bit-rate range.

[4] Oren Rippel, Sanjay Nair, Carissa Lew, Steve Branson, Alexander G. Anderson, and Lubomir Bourdev. Learned video compression. In ICCV, October 2019. 1, 2, 3, 6

Learned Video Compression

☐ Deep schemes for low-latency scenarios： Liu_AAAI2020 [5]

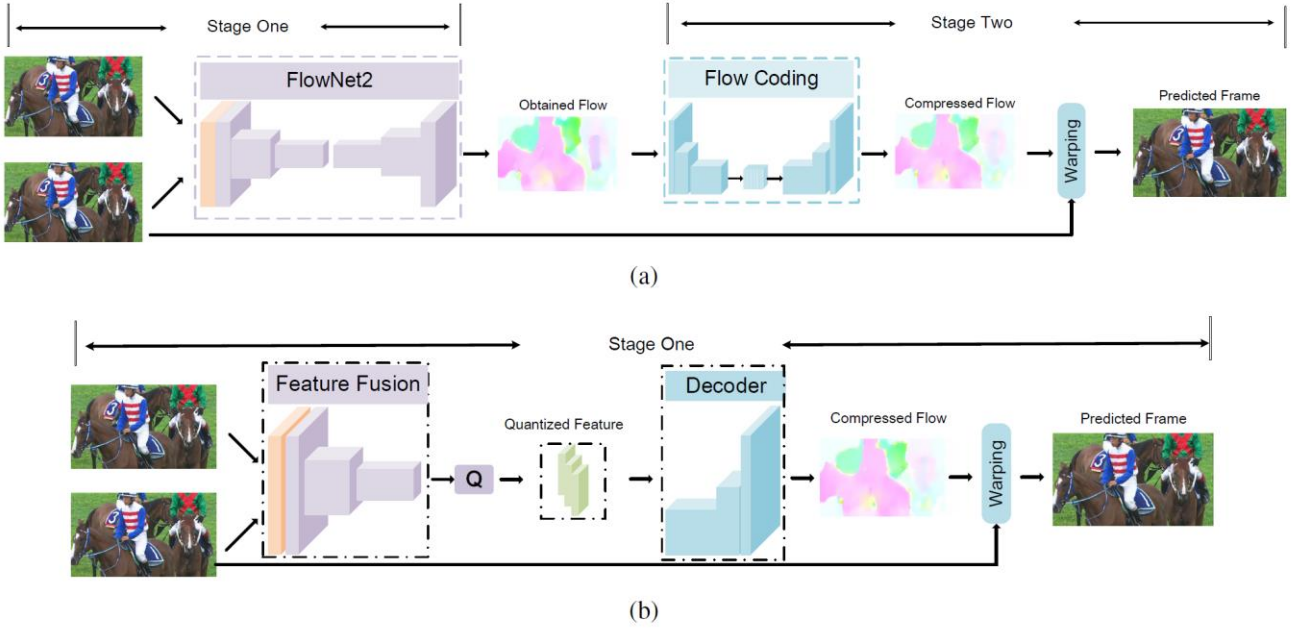


Figure 4: **Flow Learning and Compensation.** (a) Two-stage supervised approach using a pre-trained flow net (with explicit raw flow) and a cascaded flow compression autoencoder; (b) One-stage unsupervised approach with implicit flow represented by quantized features that will be directly decoded for compensation.

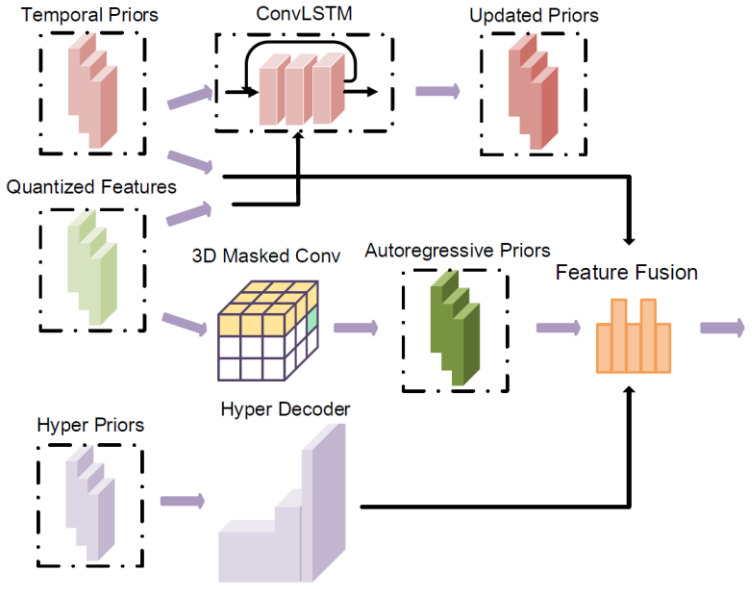
- Their model uses the one-stage flow learning and compensation, where a NLAM-based auto-encoder learns and compresses the motion information simultaneously.

- They directly use the NLAIC method proposed in their previous work to compress the residual and intra frame, where the NLA transform, hyper and aoutoregressive priors are all used.

[5] Liu, Haojie, Lichao Huang, Ming Lu, Tong Chen, and Zhan Ma. "Learned Video Compression via Joint Spatial-Temporal Correlation Exploration." arXiv preprint arXiv:1912.06348 (2019).

Learned Video Compression

☐ Deep schemes for low-latency scenarios： Liu_AAAI2020 [5]



- The entropy model of flow compression uses the adaptive contexts with fused priors, that is the temporal priors updated by a convLSTM, the autoregressive priors learned by 3D masked Conv and the Hyper priors.

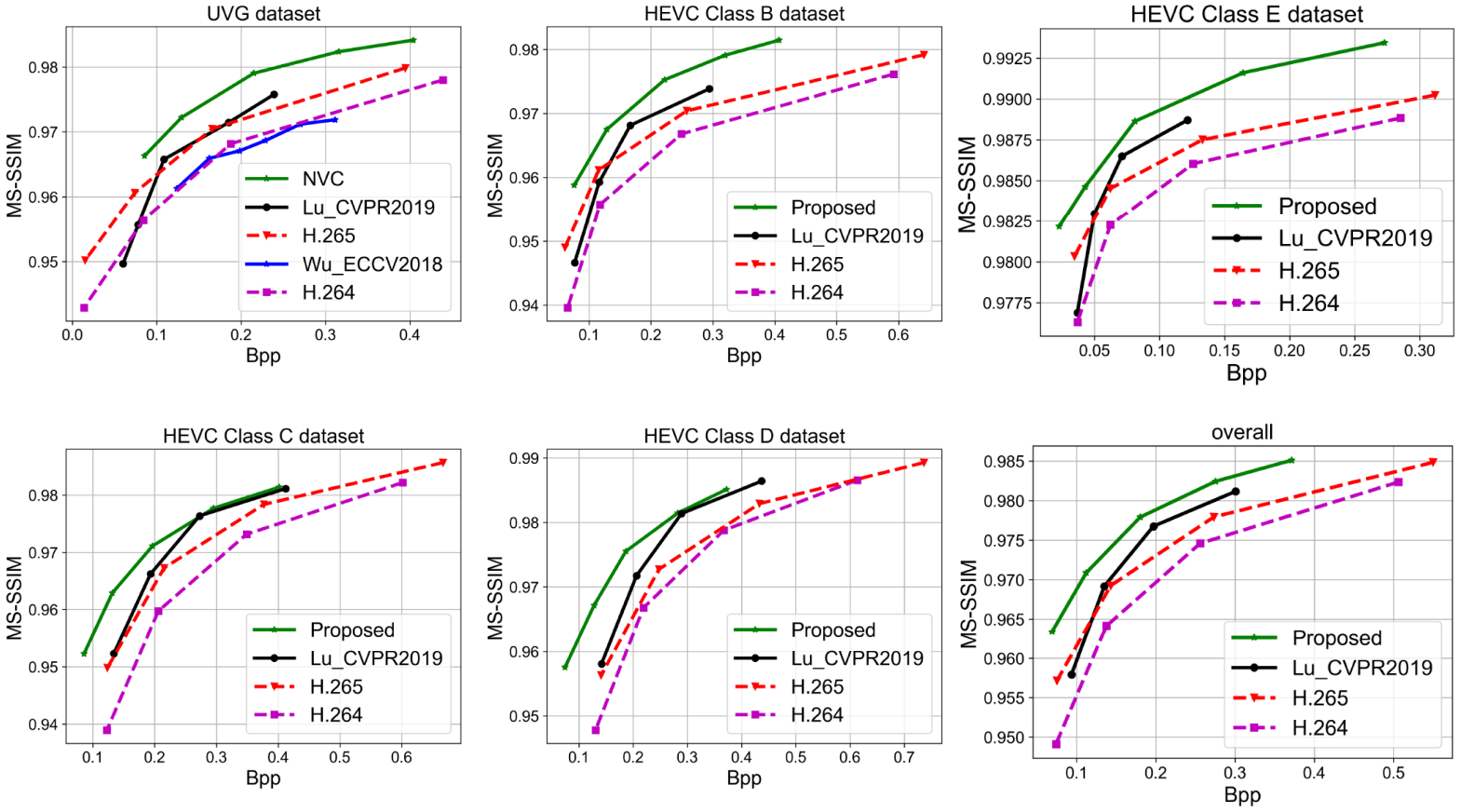
- They pre-train the intra coding and flow learning and coding networks first, followed by the jointly training with pre-trained network models for an overall optimization.

$$L = \frac{\lambda_1}{n}\sum_{t=0}^{n}\mathbb{D}_1(\hat{\mathbf{X}}_t, \mathbf{X}_t) + \frac{\lambda_2}{n}\sum_{t=0}^{n}\mathbb{D}_2(\hat{\mathbf{X}}_t^p, \mathbf{X}_t) + R_s + \frac{1}{n-1}\sum_{t=1}^{n}R_t$$

[5] Liu, Haojie, Lichao Huang, Ming Lu, Tong Chen, and Zhan Ma. "Learned Video Compression via Joint Spatial-Temporal Correlation Exploration." arXiv preprint arXiv:1912.06348 (2019).

Learned Video Compression

☐ Deep schemes for low-latency scenarios： Liu_AAAI2020 [5]

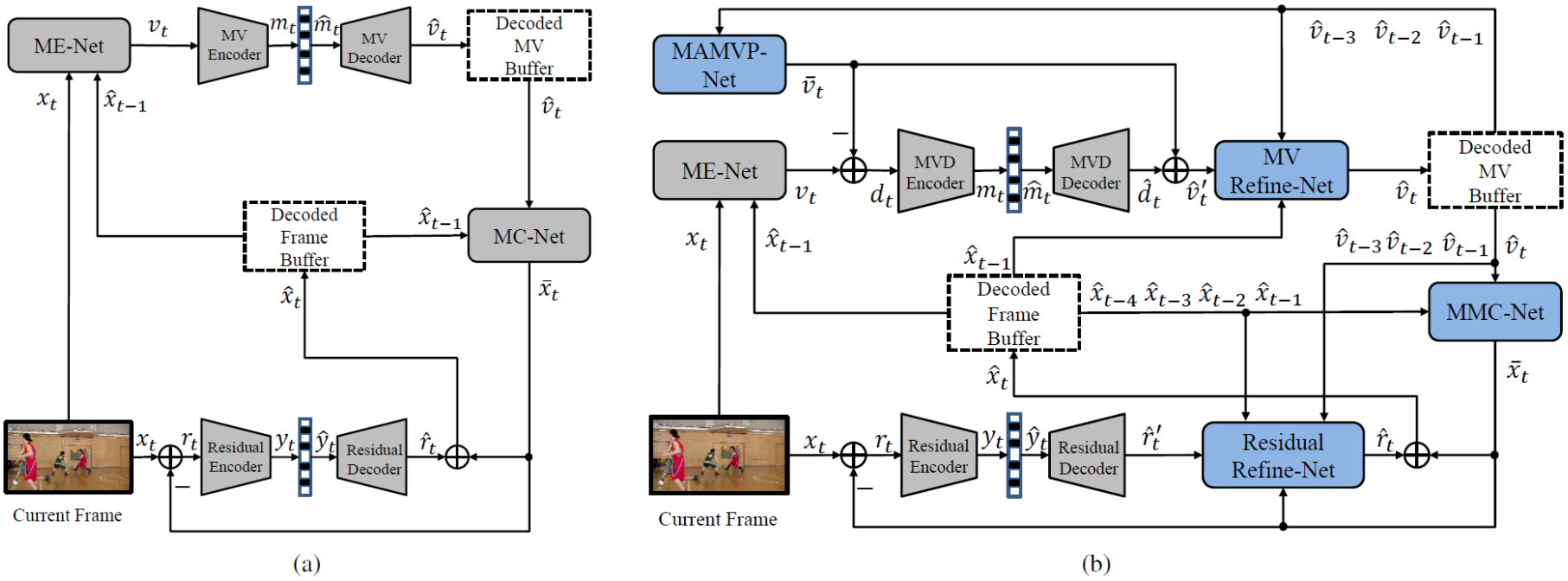


➢ From the coding results reported in the paper, we can see that their model performs better than DVC(Lu_CVPR2019) significantly. But the comparison is unfair because DVC is optimized for MSE while their method is optimized for MS-SSIM.

[5] Liu, Haojie, Lichao Huang, Ming Lu, Tong Chen, and Zhan Ma. "Learned Video Compression via Joint Spatial-Temporal Correlation Exploration." arXiv preprint arXiv:1912.06348 (2019).

Learned Video Compression

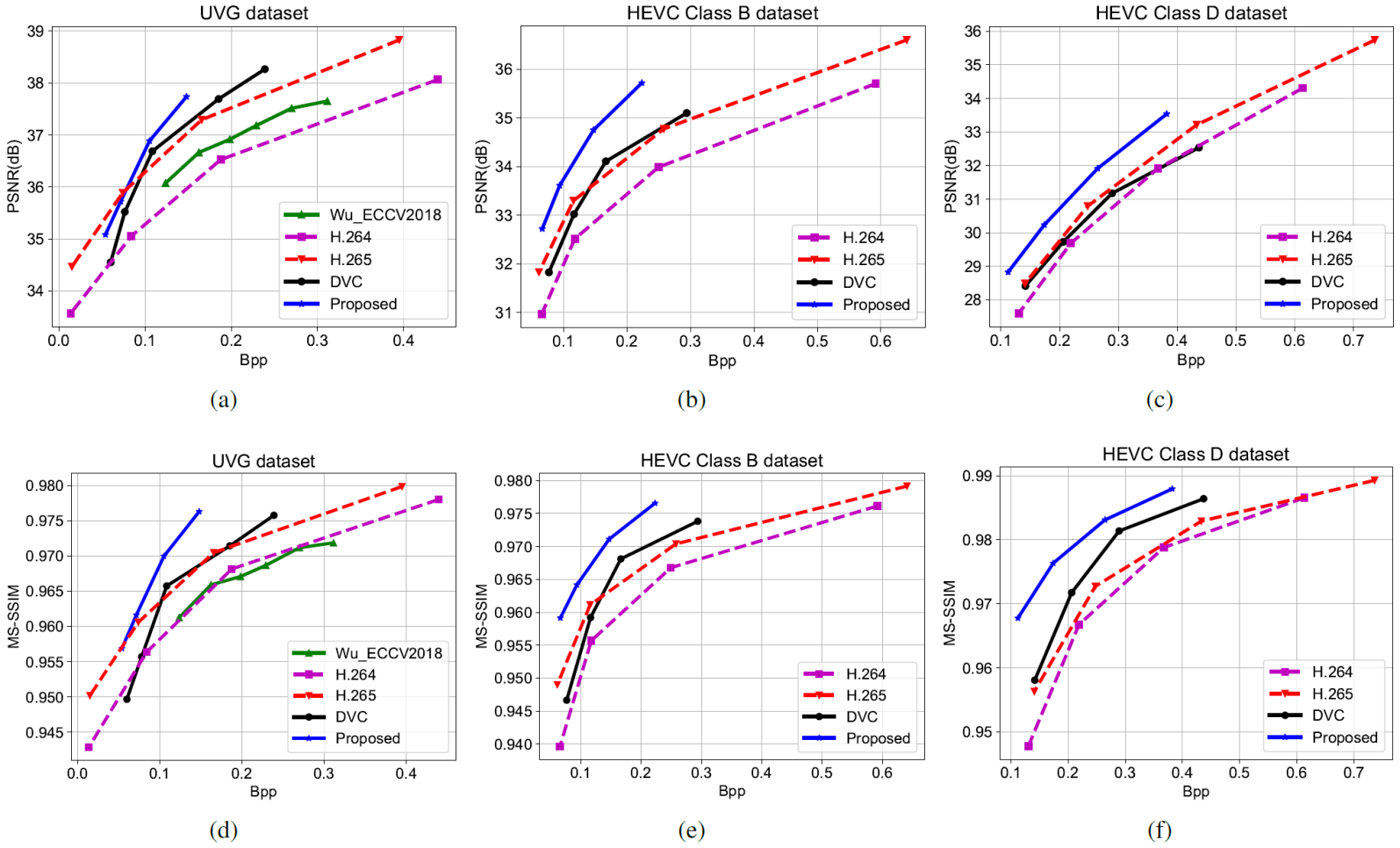
☐ Deep schemes for low-latency scenarios： Ours_CVPR2020



- Compared with DVC(a), we newly introduce four effective modules using multiple reference frames: multiple frame-based MV prediction, multiple frame-based motion compensation, MV refinement, and residual refinement.
- We use a single rate-distortion loss function, together with a step-by-step training strategy, to jointly optimize all the modules in our framework.
- Our model is optimized for MSE and uses the fully-factorized and hyperprior entropy model to compress the MVD and residual, respectively.

Learned Video Compression

☐ Deep schemes for low-latency scenarios： Ours_CVPR2020



➢ Our model performs better than DVC by a large margin in both PSNR and MS-SSIM and even performs better than Liu_AAAI2020 in MS-SSIM, although Liu_AAAI2020 is optimized for MS-SSIM and uses the auto-encoder with much higher coding efficiency.


Thank
You!