A Hybrid Framework Integrating End-to-End Learned Image Codec with Conventional Codec

Nannan Zou* ‡, Antti Hallapuro*, Francesco Cricri*, Honglei Zhang*, A. Burakhan Koyuncu†,
Jukka I. Ahonen* ‡, Miska M. Hannuksela*, Esa Rahtu‡

**Nokia Technologies*, Tampere, Finland
†*Nokia Technologies*, Munich, Germany
‡*Tampere University*, Tampere, Finland
Email: {nannan.zou, antti.hallapuro, francesco.cricri, honglei.1.zhang, burakhan.koyuncu,
jukka.1.ahonen, miska.hannuksela}@nokia.com, esa.rahtu@tuni.fi

Abstract—**This paper proposes a hybrid framework that integrates an end-to-end learned image codec (LIC) with a conventional video codec (CVC). The framework involves using LIC-coded intra frames and CVC-coded inter frames. For each intra frame, the encoder decides whether to code it with LIC or with CVC. To further enhance perceptual performance of this framework, the LIC model is additionally finetuned using perceptual quality objectives. The experimental results show that, compared to NNVC-7.1 VTM (with neural network tools off), the proposed hybrid method achieves average Bjontegaard Delta (BD)-rate savings of -0.83%, -2.03%, and -1.66% under Random Access (RA) configurations for the Y, U and V components, respectively, when PSNR is used as the quality metric. Furthermore, the method provides an average BD-rate savings of -10.06% under RA configurations for the Y component when evaluated with perceptual quality metrics.**

Index Terms—**Hybrid video coding, learned image codec, conventional video codec, perceptual quality**

I. Introduction

Video coding has attracted growing attention due to the rising demand for high-resolution, dynamic video content driven by the rapid development of digital media. Conventional video codecs (CVCs) have served as the backbone of video compression for decades. These codecs are built on handcrafted frameworks that leverage both temporal and spatial redundancy in video signals through a pipeline of motion estimation, transform coding, quantization, and entropy coding. The prominent standards in this family include H.264/AVC [1], H.265/HEVC [2], and the more recent H.266/VVC [3]. However, CVCs are constrained by their fixed pipeline and heuristic design choices, limiting their ability to fully exploit the statistical structure of video content. As a result, they are increasingly being challenged by learned video codecs, which aim to learn optimal representations directly from data. Nonetheless, CVCs continue to dominate real-world deployments due to their superior video compression efficiency, broad application support, and low decoding complexity.

In recent years, end-to-end (E2E) learned image codecs (LICs) [4]–[6] have emerged as powerful compression approaches, utilizing deep neural networks (NNs) to jointly optimize all components of the coding pipeline. These methods typically employ an autoencoder architecture, where the encoder compresses the image into a compact latent representation, which is then quantized and entropy-coded. The corresponding decoder reconstructs the image from the compressed latent, with the entire system trained end-to-end using a rate-distortion objective [7]. To further improve perceptual quality of LICs, recent works have investigated perceptual loss functions [8], adversarial training [9], and NN-based metrics [10] to align the reconstruction quality more closely with human visual perception. E2E learned image codecs have demonstrated superior performance over state-of-the-art conventional codecs such as VVC [11]. Nevertheless, in the domain of video coding, E2E learned video codecs have yet to match the performance of handcrafted solutions like VVC, particularly under the Random Access configurations.

The flexibility and effectiveness of NN-based components have motivated exploration into their integration with classical codecs such as HEVC and VVC. In [12], NN-based frame prediction was introduced to improve the coding performance of HEVC. Similarly, [13] proposed the incorporation of a lossless image codec alongside a restoration network within a conventional video codec, where the restoration network enhance the decoded frame by using guidance from the losslessly coded reference frame at the decoder. In [14], a hybrid video codec integrating a self-supervised LIC with standard VVC was presented, targeting machine-centric video analysis. NN-based filters represent another promising direction for improving the coding efficiency of CVCs. Specifically, NN-based in-loop filters were designed in [15]–[17] to enhance the reference pictures for subsequent coding, while NN-based post-processing filters were proposed in [18]–[20] to improve the quality of reconstructed output frames.

This paper presents a hybrid framework that exploits the high performance of LICs while retaining the advantages of CVCs for efficient inter-frame coding. In this framework, reconstructed LIC-coded intra frames are utilized as reference pictures for inter-frame coding performed by a CVC. Specifically, intra frames are temporarily encoded by both the LIC and the standard CVC encoder, and the encoding method with lower rate-distortion cost is chosen. The decoder will invoke either the LIC intra decoding or the CVC intra decoding,

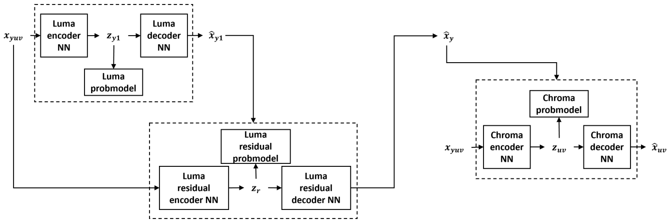


Fig. 1. The architecture of our LIC.

depending on the decision made by the encoder. The resulting reconstructed intra frame is then used as a reference for inter-frame decoding. To further improve visual quality, the LIC is additionally finetuned using perceptual losses. The approach described here operates at the granularity of the entire fame.

II. Proposed Method

A. Learned Image Codec (LIC)

Our proposed method uses an E2E learned image codec (LIC) as the intra frame codec. The LIC is designed for YUV 4:2:0 format, in which the Luminance (Y) and Chrominance (U, V) components of input frames are processed separately. More specifically, this codec comprises NN-based encoders, NN-based decoders, NN-based probability models, and an Asymmetric Numeral Systems (ANS) entropy codec [21]. Fig. 1 shows the high-level architecture. For the Luminance, the whole input frame x_{yuv} is transformed by a Luma encoder E_{y1} (parametrized by $\theta_{E_{y1}}$), obtaining the latent representation $z_{y1} = E_{y1}(x_{yuv}; \theta_{E_{y1}})$. This latent is then quantized and entropy coded by the arithmetic encoder. At decoder side, the luminance bitstream is entropy decoded to the quantized latent tensor $\hat{z}_{y1}$ and then passed through a Luma decoder D_{y1} (parametrized by $\theta_{D_{y1}}$) to reconstruct the initial luminance reconstruction $\hat{x}_{y1} = D_{y1}(\hat{z}_{y1}; \theta_{D_{y1}})$ via dequantization and decoding. To further enhance reconstruction fidelity, the residual error between the original input frame x_{yuv} and the initial luminance reconstruction $\hat{x}_{y1}$ is encoded by a Luma residual encoder E_r. The resulting residual latent z_r is modeled by a Luma residual probability model for entropy estimation and then entropy compressed. The residual bitstream is entropy decoded and subsequently decompressed by a Luma residual decoder D_r, which refines the initial reconstruction $\hat{x}_{y1}$ to generate the final reconstructed luminance output $\hat{x}_y$.

For the Chrominance, a similar but simplified process is employed. The input frame x_{yuv} is passed through a Chroma encoder E_{uv} to generate a latent representation z_{uv}, which is then quantized and entropy coded. Next, the entropy decoder decompresses the resulting chrominance bitstream, and a Chroma decoder D_{uv} is applied to reconstruct the decoded bitstream back to the pixel domain $\hat{x}_{uv}$. Notably, the final luminance output $\hat{x}_y$ is used as a conditioning signal during the chrominance coding process to enhance reconstruction quality.

The training process is organized in two stages. For both stages, the overall rate loss is defined as:

$$\mathcal{L}_{\text{rate}} = \mathcal{L}_{\text{y1}} + \mathcal{L}_{\text{r}} + \mathcal{L}_{\text{uv}} \tag{1}$$

where $\mathcal{L}_{\text{y1}}$, $\mathcal{L}_{\text{r}}$, and $\mathcal{L}_{\text{uv}}$ represent the rate losses associated with the luminance encoding, luminance residual encoding, and chrominance encoding, respectively. In the first stage, the LIC is optimized using Mean-Squared Error (MSE) and rate loss as the training objective:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{mse}} + w_{\text{rate}}\mathcal{L}_{\text{rate}} \tag{2}$$

In the second stage, the model is further finetuned to enhance perceptual quality by incorporating perceptual loss into the objective. Specifically, the total loss function is defined as:

$$\mathcal{L}_{\text{total}} = w_{\text{mse}}\mathcal{L}_{\text{mse}} + w_{\text{perceptual}}\mathcal{L}_{\text{perceptual}} + w_{\text{rate}}\mathcal{L}_{\text{rate}} \tag{3}$$

In Equation (2) and Equation (3), w_{mse}, $w_{\text{perceptual}}$, and w_{rate} denote hyper-parameters empirically set to align the rate losses with those of CVC intra coding.

Algorithm 1 Hybrid Video Coding

```
if PerceptualMode is False then
    Encode $x_{yuv}$ using CVC $\rightarrow$ $bitstream_{cvc}$, $RD_{cvc}$
    Encode $x_{yuv}$ using LIC $\rightarrow$ $bitstream_{lic}$, $RD_{lic}$
    if $RD_{lic} < RD_{cvc}$ then
        Add $\hat{x}_{yuv\_lic}$ to CVC's reference picture buffer
        Add $bitstream_{lic}$ to video bitstream
    else
        Use $\hat{x}_{yuv\_cvc}$ as CVC's reference picture
        Add $bitstream_{cvc}$ to video bitstream
    end if
else
    Encode $x_{yuv}$ using LIC $\rightarrow$ $bitstream_{lic}$
    Add $\hat{x}_{yuv\_lic}$ to CVC's reference picture buffer
    Add $bitstream_{lic}$ to video bitstream
end if
```

B. Hybrid Video Coding Framework

Rate-distortion optimization (RDO) is a process where the encoder evaluates different coding options for a given input to determine the best trade-off between bitrate and quality. The proposed hybrid video coding framework pipeline is detailed in Algorithm 1. Firstly, the input intra frame x_{yuv} is temporarily encoded using both the LIC and the standard CVC encoder. Next, the encoder decides whether to use the standard CVC-coded frame or the LIC-coded frame based on the CVC's RDO decision J, which is formulated by distortion D and bitrate R.

$$J = D + \lambda \cdot R \tag{4}$$

where λ, derived from QP, denotes the Lagrange parameter, and D indicates MSE in our implementation. If the encoder decides to use the LIC-coded intra frame, the decoded frame from the LIC is added to CVC's decoded picture buffer and is used as a reference frame for inter-frame prediction. Additionally, the LIC bitstream for the coded intra frame is included in the video bitstream. Conversely, if the standard CVC-coded frame is selected, the LIC bitstream for the coded intra frame is not included in the video bitstream. Importantly, for the case where the LIC was finetuned by using a perceptual training loss, the input intra frame is always encoded by the LIC in the simulation.

Additionally, a new Network Abstraction Layer (NAL) unit type is introduced to indicate and embed the LIC coded intra frames in the video bitstream. NN-coded intra picture header is introduced to transfer data for LIC encoded picture.

C. *Quality Metrics*

We adopt Peak Signal-to-Noise Ratio (PSNR) along with 8 perceptual quality metrics to comprehensively evaluate the performance improvement brought by the proposed hybrid framework. The selected perceptual quality metrics include: Multi-Scale Structural Similarity Index (MS-SSIM) [22], Visual Information Fidelity (VIF) [23], Feature Similarity Index (FSIM) [24], Normalized Laplacian Pyramid Distance (NLPD) [25], Information-Weighted Structural Similarity Index (IW-SSIM) [26], Video Multi-method Assessment Fusion (VMAF) [27], Peak Signal-to-Noise Ratio – Human Visual System (PSNR-HVS) [28], and Learned Perceptual Image Patch Similarity (LPIPS) [10]. To clarify, all perceptual quality metrics are computed using only the luminance (Y) component of the video frames, and VMAF is computed on a frame-by-frame basis rather than over the whole sequence.

III. Experiments

A. *Experimental Settings*

Evaluation configuration: We evaluate the proposed hybrid framework on the Common Test Conditions (CTC) sequences [29] recommended by JVET. The sequences and their corresponding resolutions are listed in Table I. In our experiments, NNVC-7.1 VTM (NN tools off) serves as the anchor and our proposed method is also built on top of it. Notably, our approach is compatible with other CVCs. All evaluations are implemented against five QPs (22, 27, 32, 37, 42) under All-Intra (AI) and Random-Access (RA) configurations.

Training strategy: We train our LIC models on 256x256 patches from the LSDIR dataset [30] which contains 84,991 high-resolution images. The images are converted to YUV 4:2:0 format with 8-bit and 10-bit depths. The resulting YUV dataset is then used for model training. A total of 5 LIC models are trained, each corresponding to respective bitrate of VTM intra coding when sequence QPs are 22, 27, 32, 37, 42.

In the first training stage, the LIC is trained for 550 epochs with a batch size of 36 using the Adam optimizer. The initial learning rate is set to 5×10^{-5} and decayed by a factor of 0.1 at the 450th and 520th epochs. During the second training stage, the perceptual loss is defined as $1 - \text{MS-SSIM}$. Following the initial training phase, finetuning is performed for additional 150 epochs with a reduced learning rate of 1×10^{-5}. Hyper-parameters w_{mse}, $w_{\text{perceptual}}$, and w_{rate} are empirically determined to align the bitrate with those of VTM intra coding.

B. *Experimental Results*

To evaluate the coding performance of our proposed hybrid framework without perceptual finetuning (W/O PFT) but with the switching mechanism enabled, we adopt PSNR as the quality metric, consistent with the distortion metric (MSE) used for switching decisions. Table I presents the results on the test sequences under both AI and RA configurations when compared with NNVC-7.1 VTM (NN tools off). Notably, in the reported AI results, input intra frames are same as in the RA configurations (i.e., they are a subset of the intra frames usually used in AI configurations). And we can observe that our proposed hybrid framework outperforms NNVC-7.1 VTM (NN tools off) on average Bjontegaard Delta (BD)-rate performance. Specifically, our proposed method can achieve an average of -0.83%, -2.03%, -1.66% BD-rate savings under RA configurations regarding PSNR metric on Y, U, and V, respectively. The results indicate the potential of our hybrid framework in improving coding performance. According to the NNVC CTCs, the QPs for intra frames in AI configurations differ from those used for intra frames in RA configurations. Therefore, the BD-rate gains under AI configurations are 0.00% for most test sequences, indicating that the encoder predominantly selects VTM-coded intra frames. This behavior is expected, as our LIC models are trained to harmonize with the bitrate levels of VTM intra coding under RA configurations.

In contrast, for the system with perceptual finetuning (W/ PFT), the switching mechanism is disabled in order to highlight the improvements in perceptual quality. Table II shows the corresponding PSNR results for the test sequences under both AI and RA configurations, compared against NNVC-7.1 VTM (NN tools off). As presented in the table, the W/ PFT system generally underperforms the anchor in terms of PSNR, particularly on the luminance component. This outcome is reasonable, as the LIC model is optimized to enhance perceptual quality rather than PSNR. The performance of both the W/O PFT and W/ PFT systems across all perceptual quality metrics is summarized in Table III. While the perceptual quality of the W/O PFT system benefits from the switching mechnism, the W/ PFT system consistently achieves significant improvements across all metrics for both AI and RA configurations. These results demonstrate a reliable and stable improvement in perceptual quality provided by the finetuned LIC.

C. *Complexity Analysis*

Table V reports the complexity of the LIC model in terms of the model size and the number of million multiply-accumulate operations per pixel (MMACs/pixel). Additionally, Table IV presents the average decoding-time measurements of intra frames for VTM and LIC, respectively. For the LIC model, decoding time is measured on both CPU and GPU across low and high coding quality settings. The measurements were respectively obtained on a single Intel® Xeon® E5-2698 v4 @ 2.20 GHz CPU and an A100 GPU. The results show that LIC achieves lower decoding time on GPU but incurs higher decoding time on CPU compared to VTM across all sequence classes, because the LIC model has not been optimized in terms of complexity.

D. *Visual Examples*

In this subsection, we provide a few examples in order to assess the impact of the LIC (both W/O PFT and W/ PFT) visually other than by means of objective metrics. In

TABLE I
PSNR-based Performance of Proposed hybrid framework W/O PFT (Switch) Compared with NNVC-7.1 VTM (NN tools off)

Class	Sequence	Resolution	All-Intra configurations			Random-Access configurations		
			Y-PSNR	U-PSNR	V-PSNR	Y-PSNR	U-PSNR	V-PSNR
A1	Tango2	3840x2160	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
	FoodMarket4	3840x2160	0.00%	0.00%	0.00%	-1.12%	-1.20%	-0.86%
	Campfire	3840x2160	0.00%	0.00%	0.00%	-2.12%	1.43%	-1.97%
A2	CatRobot	3840x2160	0.00%	0.00%	0.00%	-0.33%	-1.06%	-0.59%
	DaylightRoad2	3840x2160	0.00%	0.00%	0.00%	0.11%	0.18%	0.44%
	ParkRunning3	3840x2160	0.00%	0.00%	0.00%	-2.65%	1.08%	3.97%
B	MarketPlace	1920x1080	0.00%	0.00%	0.00%	0.77%	-0.98%	0.61%
	RitualDance	1920x1080	0.00%	0.00%	0.00%	-0.42%	-2.62%	-0.95%
	Cactus	1920x1080	0.00%	0.00%	0.00%	-1.52%	-8.32%	-3.73%
	BasketballDrive	1920x1080	0.00%	0.00%	0.00%	-0.05%	-0.27%	-0.13%
	BQTerrace	1920x1080	0.00%	0.00%	0.00%	-1.87%	-3.10%	-3.57%
C	BasketballDrill	832x480	0.00%	0.00%	0.00%	-1.15%	-5.01%	-3.90%
	BQMall	832x480	0.00%	0.00%	0.00%	-0.89%	-4.62%	-4.76%
	PartyScene	832x480	0.00%	0.00%	0.00%	-0.23%	-0.86%	-0.71%
	RaceHorses	832x480	0.00%	0.00%	0.00%	-0.60%	-2.98%	-5.10%
D	BasketballPass	416x240	0.00%	0.00%	0.00%	-0.54%	-3.94%	-1.29%
	BQSquare	416x240	0.00%	0.00%	0.00%	-0.95%	-0.68%	-2.23%
	BlowingBubbles	416x240	0.00%	0.00%	0.00%	-0.33%	-1.40%	-0.19%
	RaceHorses	416x240	-0.43%	-0.32%	-0.51%	-1.86%	-4.19%	-6.57%
Average			**-0.02%**	**-0.02%**	**-0.03%**	**-0.83%**	**-2.03%**	**-1.66%**

TABLE II
PSNR-based Performance of Proposed hybrid framework W/ PFT (NO Switch) Compared with NNVC-7.1 VTM (NN tools off)

Class	Sequence	Resolution	All-Intra configurations			Random-Access configurations		
			Y-PSNR	U-PSNR	V-PSNR	Y-PSNR	U-PSNR	V-PSNR
A1	Tango2	3840x2160	17.07%	119.34%	66.39%	5.69%	7.03%	4.29%
	FoodMarket4	3840x2160	3.77%	37.54%	33.44%	2.35%	-2.12%	-1.95%
	Campfire	3840x2160	21.50%	221.65%	37.06%	-2.91%	8.46%	-3.58%
A2	CatRobot	3840x2160	15.28%	22.38%	39.57%	-0.28%	-5.09%	0.46%
	DaylightRoad2	3840x2160	35.34%	38.19%	93.75%	2.35%	-11.07%	1.19%
	ParkRunning3	3840x2160	-14.82%	84.53%	124.42%	-3.22%	2.62%	4.76%
B	MarketPlace	1920x1080	4.14%	21.51%	11.22%	-0.70%	-7.67%	-7.21%
	RitualDance	1920x1080	4.21%	26.65%	4.86%	-0.41%	-1.57%	-2.15%
	Cactus	1920x1080	10.53%	53.22%	52.89%	-0.43%	-6.54%	2.27%
	BasketballDrive	1920x1080	25.25%	56.59%	76.14%	0.64%	-2.97%	1.40%
	BQTerrace	1920x1080	22.39%	45.03%	17.85%	0.08%	-12.62%	-23.83%
C	BasketballDrill	832x480	3.83%	27.53%	42.90%	-0.55%	3.25%	5.43%
	BQMall	832x480	4.66%	5.74%	1.20%	0.20%	-4.28%	-5.14%
	PartyScene	832x480	11.08%	15.84%	16.77%	5.20%	4.87%	11.97%
	RaceHorses	832x480	7.17%	18.05%	5.04%	0.16%	3.13%	-6.75%
D	BasketballPass	416x240	-1.92%	6.17%	19.98%	-0.89%	0.49%	1.62%
	BQSquare	416x240	8.73%	3.40%	-10.26%	1.96%	-6.58%	-11.40%
	BlowingBubbles	416x240	4.52%	12.56%	12.95%	2.34%	10.88%	15.19%
	RaceHorses	416x240	-5.13%	10.65%	5.33%	-2.60%	0.90%	-2.25%
Average			**9.35%**	**43.50%**	**34.29%**	**0.47%**	**-0.99%**	**-0.83%**

TABLE III
Perceptual Performance of Proposed hybrid framework Compared with NNVC-7.1 VTM (NN tools off) on Classes A–D

Configuration	Test	Average	MS-SSIM Torch	VIF	FSIM	NLPD	IW-SSIM	VMAF	psnrHVS	LPIPS
AI	W/O PFT (SWITCH)	-0.02%	-0.01%	-0.03%	-0.01%	-0.02%	-0.01%	-0.02%	-0.02%	-0.02%
	W/ PFT (NO SWITCH)	-16.77%	-27.67%	-1.57%	-23.21%	-12.67%	-24.48%	-13.52%	-7.18%	-14.25%
RA	W/O PFT (SWITCH)	-5.90%	-6.79%	-5.16%	-6.29%	-5.31%	-6.67%	-5.71%	-4.27%	-5.37%
	W/ PFT (No SWITCH)	-10.06%	-13.29%	-7.96%	-8.69%	-7.63%	-11.93%	-6.16%	-4.83%	-14.75%

Fig. 2, we show pictures coded with anchor, with LIC W/O PFT and with LIC W/ PFT. Both intra and inter pictures are provided. For each example picture, we provide related objective metrics as reference. As can be observed, higher values of reported perceptual metrics correlate with visually higher quality. Also, it can be observed that finetuning the LIC on a perceptual metric (MS-SSIM) improves the perceptual metrics and the visual quality. Finally, it can be observed that perceptual improvements can be propagated from intra frames to inter frames.

TABLE IV
Average decoding-time measurements of intra frames for LIC Compared with NNVC-7.1 VTM (NN tools off)

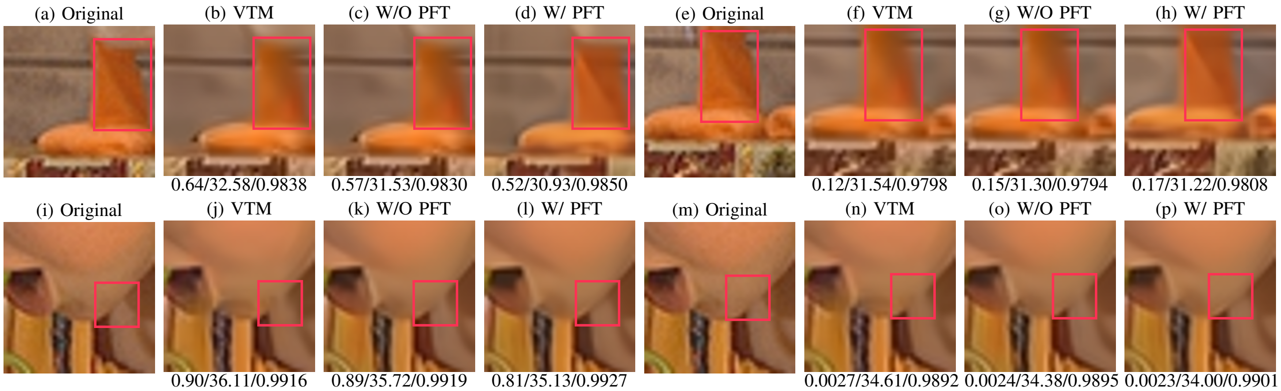
Class	Runtime (s) for QP 22					Runtime (s) for QP 42)				
	VTM	LIC (GPU)		LIC (CPU)		VTM	LIC (GPU)		LIC (CPU)	
	Abs.	Abs.	Rel.	Abs.	Rel.	Abs.	Abs.	Rel.	Abs.	Rel.
A	1.45	0.77	0.53	497.09	342.82	0.74	0.72	0.97	510.52	689.89
B	0.45	0.21	0.47	120.96	268.80	0.24	0.20	0.83	120.75	503.13
C	0.17	0.08	0.47	19.44	114.35	0.11	0.07	0.64	19.39	176.27
D	0.09	0.07	0.78	4.41	49.00	0.08	0.05	0.63	4.46	55.75
Average	**0.54**	**0.28**	**0.52**	**160.48**	**297.19**	**0.29**	**0.26**	**0.90**	**163.78**	**564.76**

IV. Conclusion

In this paper, we propose a hybrid video coding framework featuring a learned intra codec coupled with a traditional video codec, demonstrating improved rate-distortion performance. Additionally, experimental results indicate that the proposed method significantly enhances perceptual quality metrics and visual quality, particularly when the LIC is finetuend with a perceptual loss. In the future, we plan to further reduce the

TABLE V
Complexity Measurement of LIC

	MMACs/pixel	Params (M)
Encoder	1.69	32.07
Decoder	2.25	65.65
Total	3.94	97.72



(a) Original (b) VTM (c) W/O PFT (d) W/ PFT (e) Original (f) VTM (g) W/O PFT (h) W/ PFT

0.64/32.58/0.9838 0.57/31.53/0.9830 0.52/30.93/0.9850 0.12/31.54/0.9798 0.15/31.30/0.9794 0.17/31.22/0.9808

(i) Original (j) VTM (k) W/O PFT (l) W/ PFT (m) Original (n) VTM (o) W/O PFT (p) W/ PFT

0.90/36.11/0.9916 0.89/35.72/0.9919 0.81/35.13/0.9927 0.0027/34.61/0.9892 0.0024/34.38/0.9895 0.0023/34.00/0.9901

Fig. 2. Visualization of reconstructed images from the original, VVC, W/O PFT, and W/ PFT. (a)-(d) and (e)-(h) show examples from frame 320 (intra frame) and frame 352 (inter frame) of sequence *PartyScene*, with QP=37. (i)-(l) and (m)-(p) are examples from frame 64 (intra frame) and frame 65 (inter frame) of sequence *BlowingBubbles*, with QP=32. Bits-per-pixel (bpp) / PSNR-Y (dB)/ MS-SSIM-Y are provided for comparison. Zoom in for better visulization.

computational complexity of the LIC and improve the gains of the hybrid framework.

References

[1] Wiegand, T., Sullivan, G.J., Bjontegaard, G. and Luthra, A., 2003. Overview of the H. 264/AVC video coding standard. IEEE Transactions on circuits and systems for video technology, 13(7), pp.560-576.

[2] Sullivan, G.J., Ohm, J.R., Han, W.J. and Wiegand, T., 2012. Overview of the high efficiency video coding (HEVC) standard. IEEE Transactions on circuits and systems for video technology, 22(12), pp.1649-1668.

[3] Bross, B., Chen, J., Ohm, J.R., Sullivan, G.J. and Wang, Y.K., 2021. Developments in international video coding standardization after AVC, with an overview of versatile video coding (VVC). Proceedings of the IEEE, 109(9), pp.1463-1493.

[4] Duan, W., Lin, K., Jia, C., Zhang, X., Ma, S. and Gao, W., 2022, July. End-to-end image compression via attention-guided information-preserving module. In 2022 IEEE International Conference on Multimedia and Expo (ICME) (pp. 1-6). IEEE.

[5] Zou, N., Zhang, H., Cricri, F., Tavakoli, H.R., Lainema, J., Hannuksela, M., Aksu, E. and Rahtu, E., 2020, September. L 2 C–learning to learn to compress. In 2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP) (pp. 1-6). IEEE.

[6] Li, B., Liang, J. and Han, J., 2022. Variable-rate deep image compression with vision transformers. IEEE Access, 10, pp.50323-50334.

[7] Ballé, J., Laparra, V. and Simoncelli, E.P., 2016. End-to-end optimized image compression. arXiv preprint arXiv:1611.01704.

[8] Blau, Y. and Michaeli, T., 2018. The perception-distortion tradeoff. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 6228-6237).

[9] Mentzer, F., Toderici, G.D., Tschannen, M. and Agustsson, E., 2020. High-fidelity generative image compression. Advances in neural information processing systems, 33, pp.11913-11924.

[10] Zhang, R., Isola, P., Efros, A.A., Shechtman, E. and Wang, O., 2018. The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 586-595).

[11] Zhang, H., Cricri, F., Tavakoli, H.R., Aksu, E. and Hannuksela, M.M., 2022. Leveraging progressive model and overfitting for efficient learned image compression. arXiv preprint arXiv:2210.04112.

[12] Choi, H. and Bajić, I.V., 2021. Affine transformation-based deep frame prediction. IEEE Transactions on Image Processing, 30, pp.3321-3334.

[13] Rhee, H., Kim, S. and Cho, N.I., 2023. Lightweight Hybrid Video Compression Framework Using Reference-Guided Restoration Network. arXiv preprint arXiv:2303.11592.

[14] Ahonen, J.I., Le, N., Zhang, H., Hallapuro, A., Cricri, F., Tavakoli, H.R., Hannuksela, M.M. and Rahtu, E., 2023, December. NN-VVC: Versatile Video Coding boosted by self-supervisedly learned image coding for machines. In 2023 IEEE International Symposium on Multimedia (ISM) (pp. 10-19). IEEE.

[15] Li, Y., Zhang, L. and Zhang, K., 2023. iDAM: Iteratively trained deep in-loop filter with adaptive model selection. ACM Transactions on Multimedia Computing, Communications and Applications, 19(1s), pp.1-22.

[16] Huang, Z., Sun, J., Guo, X. and Shang, M., 2021. Adaptive deep reinforcement learning-based in-loop filter for VVC. IEEE Transactions on Image Processing, 30, pp.5439-5451.

[17] Jia, C., Wang, S., Zhang, X., Wang, S., Liu, J., Pu, S. and Ma, S., 2019. Content-aware convolutional neural network for in-loop filtering in high efficiency video coding. IEEE Transactions on Image Processing, 28(7), pp.3343-3356.

[18] Nasiri, F., Hamidouche, W., Morin, L., Dhollande, N. and Cocherel, G., 2021, June. Model selection CNN-based VVC quality enhancement. In 2021 Picture Coding Symposium (PCS) (pp. 1-5). IEEE.

[19] Schiopu, I. and Munteanu, A., 2022. Deep learning post-filtering using multi-head attention and multiresolution feature fusion for image and intra-video quality enhancement. Sensors, 22(4), p.1353.

[20] Lam, Y.H., Zare, A., Cricri, F., Lainema, J. and Hannuksela, M.M., 2020, October. Efficient adaptation of neural network filter for video compression. In Proceedings of the 28th ACM International Conference on Multimedia (pp. 358-366).

[21] Duda, J., Tahboub, K., Gadgil, N.J. and Delp, E.J., 2015, May. The use of asymmetric numeral systems as an accurate replacement for Huffman coding. In 2015 Picture Coding Symposium (PCS) (pp. 65-69). IEEE.

[22] Wang, Z., Simoncelli, E.P. and Bovik, A.C., 2003, November. Multiscale structural similarity for image quality assessment. In The thrity-seventh asilomar conference on signals, systems and computers, 2003 (Vol. 2, pp. 1398-1402). Ieee.

[23] Sheikh, H.R. and Bovik, A.C., 2006. Image information and visual quality. IEEE Transactions on image processing, 15(2), pp.430-444.

[24] Zhang, L., Zhang, L., Mou, X. and Zhang, D., 2011. FSIM: A feature similarity index for image quality assessment. IEEE transactions on Image Processing, 20(8), pp.2378-2386.

[25] Laparra, V., Berardino, A., Ballé, J. and Simoncelli, E.P., 2017. Perceptually optimized image rendering. Journal of the Optical Society of America A, 34(9), pp.1511-1525.

[26] Wang, Z. and Li, Q., 2010. Information content weighting for perceptual image quality assessment. IEEE Transactions on image processing, 20(5), pp.1185-1198.

[27] Li, Z., Aaron, A., Katsavounidis, I., Moorthy, A. and Manohara, M., 2016. Toward a practical perceptual video quality metric. Netflix Tech Blog (2016) [online]

[28] Egiazarian, K., Astola, J., Ponomarenko, N., Lukin, V., Battisti, F. and Carli, M., 2006, January. New full-reference quality metrics based on HVS. In Proceedings of the second international workshop on video processing and quality metrics (Vol. 4, p. 4).

[29] Boyce, J., Suehring, K., Li, X. and Seregin, V., 2018. JVET common test conditions and software reference configurations. Document JVET-J1010.

[30] Li, Y., Zhang, K., Liang, J., Cao, J., Liu, C., Gong, R., Zhang, Y., Tang, H., Liu, Y., Demandolx, D. and Ranjan, R., 2023. Lsdir: A large scale dataset for image restoration. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 1775-1787).