Emerging Advances in Learned Video Compression: Models, Systems and Beyond

Chuanmin Jia[*,1] , **Feng Ye**[1] , **Siwei Ma**[2] , **Wen Gao**[2] , **Huifang Sun**[2,3] and **Leonardo Chiariglione**[2,3]

[1]Wangxuan Institute of Computer Technology, State Key Laboratory of Multimedia Information Processing, Peking University
[2]Peng Cheng Lab
[3]MPAI
cmjia@pku.edu.cn

Abstract

Video compression is a fundamental topic in the visual intelligence, bridging visual signal sensing/capturing and high-level visual analytics. The broad success of artificial intelligence (AI) technology has enriched the horizon of video compression into novel paradigms by leveraging end-to-end optimized neural models. In this survey, we first provide a comprehensive and systematic overview of recent literature on end-to-end optimized learned video coding, covering the spectrum of pioneering efforts in both uni-directional and bi-directional prediction based compression model designation. We further delve into the optimization techniques employed in learned video compression (LVC), emphasizing their technical innovations, advantages. Some standardization progress is also reported. Furthermore, we investigate the system design and hardware implementation challenges of the LVC inclusively. Finally, we present the extensive simulation results to demonstrate the superior compression performance of LVC models, addressing the question that why learned codecs and AI-based video technology would have with broad impact on future visual intelligence research.

1 Introduction

Video data has emerged as the largest category of big data in visual information processing, with its volume continuing to increase exponentially. The majority of global network traffic has been dominated by video applications with the growing number of Internet users. Consequently, advancements in image and video compression technologies have become critical, as they directly impact over half of the world's population. Driven by emerging functional requirements and application scenarios like interactive video conferencing and remote screen sharing, there is an increasing need for advanced coding techniques to ensure efficient video transmission, storage, and enhanced quality-of-experiences. Lossy video compression plays a fundamental role in handling these demands.

Numerous video compression approaches have been investigated, enabling plenty of video services and standardized contributions [Bross *et al.*, 2021; Ma *et al.*, 2022; Ma *et al.*, 2023]. More recently, deep neural networks have significantly advanced various fields, and their application to video coding has garnered increasing attention for their remarkable non-linear modeling ability to enrich the horizon of image and video coding [Ma *et al.*, 2019; Liu *et al.*, 2020]. Evidence could be observed that they achieve promising rate-distortion (R-D) performances on commonly adopted test benchmarks and realize comparable performances with the state-of-the-art standardized coding methods. Given its substantial R-D performance, flexible technological development, plug-and-play hardware deployment, and adaptability to future requirements in intelligent media applications, LVC promises to be the next major advance in video compression.

Apart from the academic research on LVC, efforts are also made from the standardization perspective. The Moving Picture Experts Group (MPEG) and Joint Video Expert Team (JVET) have been continuously studying the requirements and technological tools for neural coding solutions since 2019. Modular neural network-based coding tools provide sufficient coding gain for almost every single module of the hybrid framework, resulting in LVC by replacing existing tools with neural approaches. The Moving Picture, Audio and Data Coding by Artificial Intelligence (MPAI) organization has established a novel project named end-to-end video coding (EEV) in early 2022, targeting at longer-term needs for fully neural-network-based video compression. The Audio and Video coding Standard Group (AVS) has also initiated End-to-end video coding Exploration Model (EEM) to explore low latency and computational complexity compression models.

This paper aims to provide a *comprehensive, systematic, and up-to-date survey* on LVC, ranging from model formulation to merging methods and applications. It is believed that the survey can serve as an innovative starting point for understanding the development situation of LVC field, identifying the limitations of existing methods, and exploring potential future directions. To this end, we begin with providing an overview of the existing lossy LVC model architectures and their evolution history, which serve as the backbones of this survey in Section 2. Section 3 discusses key optimization techniques relevant to LVC, while Section 4 elaborates
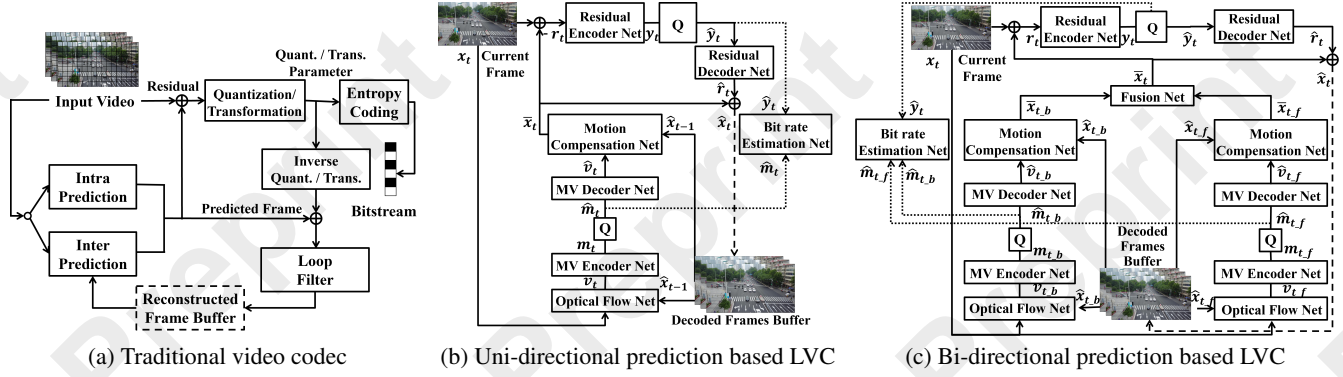
---

[*]Corresponding author.

Figure 1: The direct architecture comparisons among different video compression methods. (a) Conventional hybrid video codec, (b) Uni-directional prediction based LVC. (c) Bi-directional prediction based LVC. Conventional codec suffers from local-modular optimization with marginal efficiency improvement. The latter two structures support end-to-end (E2E) optimization.

on the system design and hardware implementation. In Section 5, we conduct experimental evaluations to demonstrate the compression performance of state-of-the-art LVC models. Finally, we present our conclusions in section 6. We deeply hope this survey provides novel insights for promoting the advanced LVC model both for the machine learning and signal processing community.

2 Overview of LVC

As depicted in Fig.1a, existing traditional video coding standards employ a block-based transform-prediction hybrid coding framework. Frames are partitioned into variable-sized blocks, with each encoded using optimal intra-prediction modes based on reconstructed neighboring blocks (left and above) for spatial context. Inter-frame prediction further enhances efficiency through motion estimation and compensation, identifying similar blocks as hypotheses. Such an approach is inherently block-dependent and requires sequential processing, which limits its flexibility and efficiency.

In contrast, LVC leveraging neural networks exhibits superior nonlinear modeling capabilities and enables end-to-end joint optimization of all modules, thereby achieving significantly higher overall performance. This advantage allows LVC frameworks to incorporate more sophisticated modules and intricate designs as shown in Fig.1b and 1c. Pioneering research in LVC, such as the works presented in [Chen *et al.*, 2017; Lu *et al.*, 2019], laid the foundation for this paradigm shift. To provide a clear and structured overview of the advancements in this field, we systematically introduce the frameworks and technical evolution of typical end-to-end optimized LVC models. The discussion is organized into two categories: uni-directional and bi-directional prediction-based methods, highlighting their respective developments and contributions to the field. Summaries of uni-directional and bi-directional predicted LVC models are shown in Table 1 and 2.

2.1 Uni-Directional Prediction Based LVC
Since 2017, primary explorations on LVC have initially focused on pixel-domain operations to reduce redundancy.

Early works such as [Chen *et al.*, 2017; Lu *et al.*, 2019] utilize pixel-level optical flow for motion estimation and motion compensation to compress residuals. These methods employ complex structural designs, such as pre-trained optical flow networks and complicated training schemes. Therefore, pixel-domain operations have inherent limitations. First, current optical flow networks struggle to generate accurate pixel-level optical flow information, especially for high-resolution videos with complex motion. Second, pixel-domain operations may introduce significant artifacts during motion compensation. Furthermore, producing reliable pixel-level residual information is challenging, and compressing it is inherently difficult. Despite these limitations, these early models lay the foundation for end-to-end optimized video codecs, and their shortcomings in model design and structure organization spur rapid advancements in subsequent research.

To address the limitations of pixel-level residuals, researchers shifted their focus to feature-domain operations. In 2020, a significant advancement in this direction involves calculating residuals in high-dimensional feature space and transforming them into discrete latent representations, rather than directly computing pixel-level residuals, as demonstrated in [Feng *et al.*, 2020]. This approach marked a notable departure from traditional pixel-domain methods. Building on this progress, a feature space video coding network was developed to perform all operations including motion estimation, motion compensation, and residual compression in the feature domain, as shown in [Hu *et al.*, 2021]. By fusing multiple reference features stored in the decoding buffer to reconstruct frames, this method achieved superior compression performance. Further advancements include learning temporal contexts from propagated features rather than previously reconstructed frames, as proposed in [Sheng *et al.*, 2022], which enhanced system performance by capturing more meaningful temporal information during encoding.

Residual coding has been a cornerstone of LVC methods, particularly before 2020, due to its ability to effectively leverage the strong temporal correlations between frames. By encoding the difference between the current frame and its prediction from reference frames, residual coding has proven to

Methodology	Motion Estimation	Motion Compensation	Coding Manner	Feature Extractor	MRF	VBSM	MMC
DVC[2019]	✓	Pixel	Residual	✗	✗	✗	✗
RaFC [2020]	✓	Pixel	Residual	✗	✗	✗	✗
NVC [2021]	✓	Feature	Residual	✗	✗	✗	✓
DCVC [2021]	✓	Feature	Conditional	✓	✗	✗	✗
ELF-VC [2021]	✓	Pixel	Residual	✗	✗	✓	✗
VLVC(lowdelay) [2021]	✓	Pixel	Residual	✗	✓	✗	✗
DCVC-HEM [2022]	✓	Feature	Conditional	✓	✗	✓	✓
C2F-DVC [2022]	✓	Feature	Residual	✓	✗	✗	✓
CANF-VC [2022]	✓	Pixel	Conditional	✗	✓	✗	✗
Alpha-VC [2022]	✗	-	Residual	✓	✗	✗	✗
MPAI EEV-0.4 [2023]	✓	Feature	Residual	✓	✗	✗	✗
DCVC-FM [2024]	✓	Feature	Conditional	✓	✗	✓	✓
DCVC-SDD [2024a]	✓	Feature	Conditional	✓	✗	✓	✓

Table 1: A comprehensive model summary of uni-directional predicted (P-frame based) LVC models. The modulized feature and network design are comparatively illustrated to better understand the design philosophy behind those models. It should be noted that the "VBSM" refers to Variable Bitrate in a Single Model, "MMC" refers to Multi-scale Motion Compensation. "MRF" refers to Multiple Reference Frames/Features, respectively.

be a practical approach. However, Shannon entropy analysis reveals that the entropy of residual coding is always greater than or equal to that of conditional coding, suggesting that residual coding is not always the most efficient choice. This limitation prompted the exploration of alternative methods, leading to the development of conditional coding.

Conditional coding emerged as a more extensible and efficient alternative to residual coding, defining conditions as learnable feature-domain contexts that can incorporate rich information to aid in encoding, decoding, and entropy modeling. Initially applied to specific modules such as entropy and foreground content coding, as demonstrated in [Ladune *et al.*, 2020], conditional coding was later expanded into a comprehensive framework. This framework utilized motion estimation and motion compensation in the feature domain to generate contextual features, significantly enhancing the capacity of conditional information, as shown in [Li *et al.*, 2021]. The potential of conditional coding was further realized in [Ho *et al.*, 2022], where it was extended to motion and integrated into a fully conditional coding-based video compression framework, achieving state-of-the-art performance and highlighting its superior advantages over traditional residual coding methods.

While the previous discussion focused on single-reference-frame scenarios, the use of multiple reference frames or features has been shown to enhance prediction performance by providing richer motion information and improving motion estimation and compensation accuracy. Compared to frameworks that rely on a single reference frame, multiple reference frames can significantly reduce the residual or conditional information that needs to be encoded, thereby lowering the bit rate and improving coding efficiency. For instance, the use of multiple previous frames as references was proposed to generate more accurate predictions and minimize residual

information, accompanied by a motion vector (MV) prediction module to extrapolate and predict MVs, further reducing coding costs [Lin *et al.*, 2020]. Additionally, multiple reference frames help the encoder adapt to diverse scenes and content, such as sequences with stationary backgrounds, complex motion, or occlusion, where they can deliver substantial performance gains. A multiple feature fusion module based on a non-local attention mechanism was introduced to leverage features from multiple previous frames, aiding in reconstructing the current frame [Hu *et al.*, 2021]. This concept was further advanced by designing a flow extrapolation network to generate effective conditions for conditional motion coding, demonstrating the significant advantages of using multiple reference frames and features [Ho *et al.*, 2022].

In the rapidly evolving field of P-frame-based LVC, state-of-the-art models have demonstrated superior rate-distortion (R-D) performance compared to conventional video codecs. This progress highlights the potential of LVC for practical applications. Future research is expected to focus on improving robustness and developing content-dependent coding methods, as network structures continue to converge and mature.

2.2 Bi-Directional Prediction Based LVC

While bi-directional prediction-based methods in LVC have received less attention compared to uni-directional prediction approaches and face challenges in achieving comparable performance, they still represent a significant and evolving area of research with their own unique development trajectory and contributions. This section explores the advancements in bi-directional prediction-based LVC, focusing on key strategies like frame interpolation and hierarchical coding structures.

Bi-directional prediction leverages both preceding and subsequent reference frames, making frame interpolation an intuitive and effective method for enhancing compression. Early

Methodology	Motion coding	Frame Interpolator	Coding Manner	Feature Extractor
Compress AE [2018]	✓	✓	Residual	✗
NeuralInter [2019]	✓	✓	Residual	✗
HLVC [2020]	✓	✗	Residual	✗
LHBDC [2021]	✓	✗	Residual	✗
B-EPIC [2021]	✓	✓	Residual	✗
Conditional [2021]	✓	✗	Conditional	✗
VLVC [2021]	✓	✗	Residual	✗
B-CANF [2023b]	✓	✗	Conditional	✗
TLZMC [2023]	✗	✓	Conditional	✗
Bi/Hi-DCVC [2023]	✓	✗	Conditional	✓
DVC-SHBMM [2024]	✓	✗	Residual	✓
DCVC-B [2024b]	✓	✗	Conditional	✓

Table 2: A summary of Bi-directional predicted (B-frame based) LVC models. The modulized feature and network design are comparatively illustrated to better understand the design philosophy behind those models.

work by [Wu *et al.*, 2018] treated video compression as a repetitive image interpolation task, employing deep image generation and interpolation for hierarchical coding of B-frames. Building on this, an interpolation model was proposed by [Djelouah *et al.*, 2019] that combined motion compression and image synthesis while reducing computational overhead during decoding. Further advancements were made by [Pourreza and Cohen, 2021], who interpolated two reference frames to generate a single reference frame, which was then used alongside an existing P-frame codec to encode the input B-frame. A more recent innovation by [Alexandre *et al.*, 2023] introduced a low-resolution base layer containing interpolated frames, replacing the motion encoder to improve compression performance while reducing complexity.

In addition to frame interpolation, hierarchical coding structures have also been explored in bidirectional LVC, offering the advantage of adapting to more flexible coding configurations while enhancing compression efficiency. These structures, widely used in traditional video coding standards, allow high-quality frames to serve as superior references, improving the compression of other frames. However, research on hierarchical B-frame schemes in LVC remains limited. In [Yang *et al.*, 2020], a Hierarchical Learned Video Compression (HLVC) method with three quality layers was proposed, applying different coding methods and compression qualities to each layer. While effective, this approach suffered from a fixed layered structure and the use of separate compression models for each layer, leading to increased memory and computational costs. Addressing these limitations, [Kim *et al.*, 2023] introduced a bidirectional LVC model with a hierarchical structure similar to conventional standards, incorporating temporal layer-adaptive optimization. This approach enabled hierarchical quality regulation within a single model, offering a more efficient and flexible solution.

In 2024, significant advancements were made in bidirectional learned video compression (LVC) models. A novel bidirectional LVC model featuring two parameter-shared motion codecs was introduced in [Ye *et al.*, 2024], which effectively leverages scaled motion information from temporal contexts as prior. This model incorporates trustworthy motion modeling and an efficient rate allocation strategy to optimize

overall compression performance. DCVC-B was also developed in [Sheng *et al.*, 2024b], which utilized a bi-directional motion difference context propagation method and a contextual compression model. These developments have enabled LVC models to surpass the performance of the H.266/VVC reference software on specific test datasets under same configurations in both RGB and YUV420 color spaces.

In summary, while bidirectional prediction-based LVC methods are still in their developmental stages, they have demonstrated promising progress through innovations in frame interpolation and hierarchical coding structures. These advancements highlight the potential of bidirectional prediction to overcome current limitations and contribute to the broader evolution of learned video compression.

3 Optimization Techniques for LVC

The data-trained LVC models have different plugin-in modules and R-D behaviors compared to conventional codecs. Therefore, the optimization of LVC frameworks remains a critical area of research, as it directly impacts compression performance, adaptability, and practical applicability. This chapter delves into six key optimization methods to highlight their technical innovations, advantages, and contributions to advancing the state-of-the-art in LVC. By systematically examining these optimization strategies, this chapter provides a comprehensive understanding of how LVC frameworks can be fine-tuned to meet diverse compression requirements and application scenarios.

3.1 Encoder-Side Online Learning

LVC models tend to suffer from severe error propagation due to the accumulation of reconstructed error in low-delay interpredictive coding. To reduce the error propagation and online adapt the contents, an error propagation aware (EPA) training strategy was proposed in [Lu *et al.*, 2020]. Considered more temporal information to alleviate error accumulation, this method supported inference stage adaptation by allowing the online update of the video encoder without introducing any extra parameters. Specifically, to optimize the encoder for each frame and mitigate the domain gap between training and testing data, the online encoder updating scheme is in the inference stage. The encoder part is updated according to the input image while keeping the decoder unchanged. In contrast to other low-level vision tasks, the pristine video frames are available at the encoder side. As a result, it is practical to update the encoder by using the original frame as long as the decoder remains unchanged.

3.2 Decoder-Side Content Adaptation

Similar to conventional video codecs, the decoder side video content adaptation for quality enhancement significantly improves the reconstructed quality of decoded videos. The author borrowed the knowledge from the quality post-processing models [Yang *et al.*, 2020] to realize compression artifacts removal and noise reduction. The decoder-side motion derivation was explored in [Chen *et al.*, 2022] using the LTSM-UNet structure to efficiently capture both spatial and temporal information. In this manner, it is obvious that

the computational overhead of motion estimation and motion compensation is avoided. However, using a displacement calculation unit to store and update the motion information results in poor performances in high-bitrate coding.

3.3 Variable Bitrate Encoding

Variable bitrate encoding in LVC has been achieved through two primary approaches: network modulation and mask modeling for latent features. Network modulation techniques, such as those proposed in [Lin *et al.*, 2021], use a rate-distortion (R-D) tradeoff parameter to modulate internal feature maps of motion and residual auto-encoders, enabling variable rate adaptation within a single model. Multi-rate training strategies, including multi-rate-distortion loss functions and step-by-step optimization, have further enhanced this capability [Rippel *et al.*, 2021]. Additionally, discrete and continuous rate adaptation has been realized through mechanisms like *Gain Units* and exponential interpolation, though interpolation-based methods often face limitations in bitrate range [Cui *et al.*, 2021]. On the other hand, mask modeling for latent features, such as the content-adaptive bit allocation method in [Lin *et al.*, 2022], dynamically adjusts the quantization strategy through latent scaling. Binarized mask modeling, as explored in [Fathima *et al.*, 2023], offers a flexible and training-stable alternative for adjusting bitrates by simulating different masking ratios of latent features. These approaches collectively enable single-model multi-rate adaptation, either by implicitly smoothing entropy through auxiliary sub-modules or explicitly modeling binarized masks for latent elements.

3.4 Rate Control Models and Methods

Rate control in LVC has been addressed through optimization-based and empirical-based methods. Optimization-based approaches, such as Semi-Amortized Variational Inference (SAVI), model rate control as a GoP-level optimization problem, harmonizing parameter updates with the learning process of LVC models to achieve optimal rate control [Xu *et al.*, 2023]. Reinforcement learning techniques, including imitation learning within a Partially Observable Markov Decision Process (POMDP) framework, have also been applied to learn neural rate control policies [Mao *et al.*, 2021]. Empirical-based methods, such as the deep learning-based rate implementation network in [Zhang *et al.*, 2024b], allocate optimal bitrates to frames based on spatial and temporal characteristics. Additionally, sparse-to-dense R-D point conversion schemes, as proposed in [Chen *et al.*, 2023a], enhance rate control flexibility without increasing model complexity. While empirical methods are easier to implement, they often suffer from performance limitations, whereas optimization-based methods, despite their high computational cost, offer more precise control. The choice of rate control method ultimately depends on specific application requirements and scenarios.

3.5 Perceptual Quality Optimization

Human perceptual quality is at the core of lossy LVC, with numerous approaches developed for perceptual quality assessment and improvement. The authors of [Pergament *et al.*, 2022] developed an interactive web-tool that allows annotators to construct spatio-temporal perceptual importance maps from a video. Regarding the extremely low bitrate coding situation, the learning objective from the adversarial perspective was able to improve the perceptual quality of decoded videos [Veerabadran *et al.*, 2020]. In this context, recurrent conditional generative adversarial learning was introduced in [Yang *et al.*, 2022] to generate photo-realistic and temporally coherent compressed frames. This training strategy ensured a balance between bitrate, distortion and perceptual quality with reasonable Generalizability.

Generally speaking, the perceptual quality-oriented LVC models usually propose hybrid objective functions (signal level plus perceptual level loss) associated with enhanced learning strategies such as discriminative-generative learning, adversarial learning, region-of-interest masking, etc.

3.6 Network Model Quantization

Model quantization of neural networks is a critical step towards deployment for LVC, ensuring interoperability among different devices especially guaranteeing the entropy decoding of the bitstream. Mainstream model post-training quantization techniques for learned compression models encapsulate weights scalar quantization and calibration, to enable reliable cross-platform encoding and decoding procedures using variational models. To overcome the catastrophic mismatch of the Gaussian prior between the sending and the receiving side, [Ballé *et al.*, 2018] proposed to train the networks entirely using floating point computations, then rounded to integers after every computational operation, while the back-propagation was done with full floating point precision. In addition, they pre-computed all possible values and expressed them as a lookup table to ensure the deterministic computation of prior distribution. The range-constrained activation and the fixed-point processing for range-adaptive quantization to convolution kernels were introduced in [Hong *et al.*, 2020], keeping fixed-point arithmetic with significant complexity reduction and robust performance at inference stage.

To realize LVC interoperability, network-weights quantization is mandatory for LVC methods. However, current quantization schemes usually enjoy mix-precision bit-depth and input-dependent, which shows little friendly to the practical utility of LVC. Fixed-bit quantization of weights and activations might be harmful to the R-D efficiency, especially in high-bitrate coding scenarios. Therefore, weight calibration and quantization-aware training should be further applied after coarse quantization process.

The optimization methods discussed in this chapter underscore the versatility and adaptability of LVC frameworks. These advancements not only improve compression efficiency but also enable LVC to cater to a wide range of practical applications, from low-latency streaming to high-quality video storage. As research in LVC continues to evolve, the integration and refinement of these optimization techniques will play a pivotal role in bridging the gap between theoretical innovation and real-world deployment. The insights provided in this chapter lay the groundwork for future explorations, paving the way for more efficient, adaptive, and scalable video compression solutions.

	JCT-VC Class B	JCT-VC Class C	JCT-VC Class D	JCT-VC Class E	UVG	MCL-JCV	**Average**
DCVC-DC [2023]	-24.54	-6.12	-18.83	-62.26	-41.70	-5.23	**-26.45**
DCVC-FM [2024]	-19.27	-16.72	-25.55	-42.46	-28.81	-24.26	**-26.19**
VTM-23.0 [2021]	-40.21	-35.59	-32.15	-38.05	-39.12	-41.19	**-37.72**
EEV-0.5 [2024]	-32.10	-13.55	-29.22	-61.76	-30.21	-34.92	**-33.63**

Table 3: BD-Rate(%) performance comparison using PSNR (RGB colorspace). The intra-period is set to be 32. Baseline is HM-16.20.

	JCT-VC Class B	JCT-VC Class C	JCT-VC Class D	JCT-VC Class E	UVG	MCL-JCV	**Average**
DCVC-DC [2023]	-39.71	-43.94	-52.76	-10.68	-30.05	-33.73	**-35.15**
DCVC-FM [2024]	-5.35	-22.75	-22.99	-19.20	-3.42	-17.86	**-15.26**
VTM-23.0 [2021]	-38.02	-34.72	-31.00	-37.42	-38.77	-42.07	**-37.00**
EEV-0.5 [2024]	-45.49	-38.36	-49.67	-54.33	-33.20	-28.17	**-41.53**

Table 4: BD-Rate(%) performance comparison using MS-SSIM (RGB colorspace). The intra-period is set to be 32. Baseline is HM-16.20.

4 LVC System Design and Implementation

Towards real-world LVC utility, the system-level design, interface collaboration and hardware implementation (e.g., resource-constrained, edge-computing devices and mobile devices) are crucial parts of LVC research and significantly impact the practical usage and deployment. These aspects are particularly vital when considering the diverse hardware environments in which LVC operates, ranging from resource-constrained edge-computing devices to mobile devices, each presenting unique challenges and opportunities. In this section, we delve into system-level designs and hardware implementation strategies that underpin LVC's practical adoption.

4.1 Edge-Computing Devices Oriented LVC

The integration of hardware acceleration into LVC has emerged as a promising direction for achieving real-time and energy-efficient video coding. A novel pipeline, termed FPX-NVC (FPGA Accelerated Neural Video Coding), was proposed in [Jia *et al.*, 2022] to leverage edge-computing FPGA devices for accelerating neural operations in both intra and inter coding. The framework introduces two key contributions: an innovative network structure and an energy-efficient network deployment method. The associated intelligent video coding system was also established, composed of visual capturing, neural encoding, decoding, and display, realizing 4K ultra-high-definition (UHD) video content.

Furthermore, the deployment of learned image compression on hardware platforms, particularly field-programmable gate arrays (FPGAs), has gained significant attention due to its potential to achieve real-time processing while maintaining computational efficiency. A fine-grained pipeline architecture was proposed in [Sun *et al.*, 2022] to realize high digital signal processing efficiency, leveraging 8-bit fixed-point arithmetic for network quantization. In summary, FPGA devices accelerated NVC schemes significantly reduce the computational complexity, maintaining the model upgradeability and flexibility. Video services such as cloud-based video production and transcoding might be critical utility scenarios.

However, the disadvantages of the edge-computing acceleration for neural codecs are also obvious. The widely used neural inference engine usually supports standard operations such as convolution, softmax, ReLU-based activation, etc. Regarding the latest advanced operations, such as transformer and multi-head attention, the embedded support within those devices is limited. Therefore it is necessary to develop a more flexible LVC deployment mechanism.

4.2 Mobile Devices Oriented LVC

In the rapidly evolving field of neural video decoding, significant strides have been made to enhance real-time performance and efficiency on mobile devices. [Le *et al.*, 2022], proposed by Qualcomm, was the first inter-frame neural video decoder with an efficient network architecture running on a commercial mobile phone, decompressing 720P high-definition videos in real-time while maintaining a low bitrate and high visual quality. The special design of MobileCodec was that the separate quantization parameters for each output channel were employed for the convolutional weights, resulting in channel-wise quantization aware tuning. As its successor, the MobileNVC model [Van Rozendaal *et al.*, 2024] realized 1080P YUV420 video real-time decoding on a mobile device. The concepts of overlapped block motion compensation were enjoyed to reduce coding artifacts using the mobile neural accelerator. By doing so, the runtime memory consumption could also be reduced compared to frame-level motion compensation.

The neural inference ability on mobile phones has been exponentially increasing along with the advancement of smart phone hardware development. General-purpose neural inference engines are the building blocks and essential hardware support for LVC deployment on smartphones.

4.3 Semantic Communications Systems

Recent advancements in semantic communications have introduced innovative deep learning-based frameworks to enhance the data transmission efficiency. [Dai *et al.*, 2022] proposes Nonlinear Transform Source-Channel Coding (NTSCC), which integrates nonlinear transform coding with deep joint source-channel coding to learn latent representations and entropy models, particularly excelling in high-
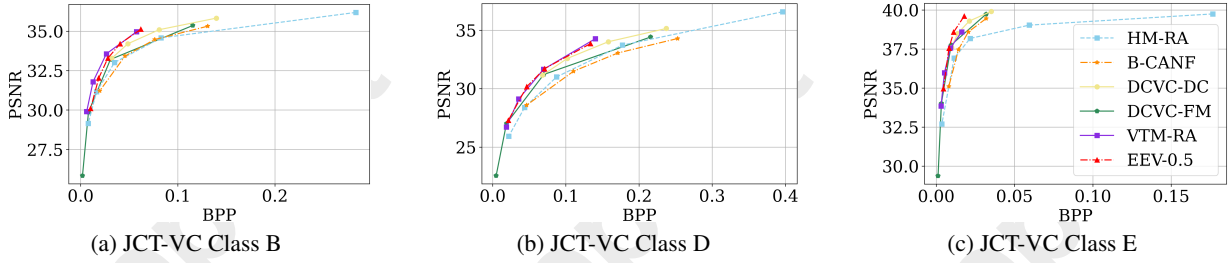
Figure 2: Rate-distortion curves for PSNR metric (RGB colorspace). HM/VTM indicates HEVC/VVC standards respectively. Four advanced LVC models are compared. (a) Class B (1920×1080). (b) Class D (416×240). (c) Class E (1280×720).

resolution image transmission. [Zhang *et al.*, 2024a] critically reviews the advancements in semantic information and semantic communications, including theory, architecture, and potential applications. Moreover, the work deeply explores key challenges in developing semantic communications and present development prospects, aiming to prompt further scientific and industrial advances in semantic communications.

5 Experiments of LVC Models

Test Datasets. Performance evaluation datasets contain JCT-VC test sequences[Boyce *et al.*, 2018], UVG dataset[Mercat *et al.*, 2020] and MCL-JCV dataset[Wang *et al.*, 2016], which are widely adopted in video compression research to measure the performance of coding models and algorithms. These test sequences include a diverse set of video clips with varying resolutions, frame rates, and content types, ensuring comprehensive testing under different scenarios.

Performance Evaluation Metrics. PSNR and MS-SSIM are two widely used metrics for evaluating video quality. PSNR measures the logarithmic ratio of the maximum pixel intensity to the Mean Squared Error (MSE) between original and compressed frames. MS-SSIM is a perceptually motivated metric that aligns more closely with the human subjective perception. BD-rate is widely used in video compression to evaluate the efficiency of coding algorithms. It quantifies the average bitrate savings or overhead of one method compared to another while maintaining the same visual quality. A negative BD-rate indicates bitrate savings, meaning the evaluated method achieves the same quality at a lower bitrate, while a positive BD-rate suggests higher bitrate requirements.

Test Conditions. We evaluate our coding performances using the first 96 frames of each sequence. Our benchmarks include traditional video coding standards H.265/HEVC and H.266/VVC with random access configuration. Their reference model HM-16.20 and VTM-23.0 are selected. Regarding HM, we employ the default *encoder_randomaccess_main_rext* configuration with intra-period=32 and GOP-size=16. As for VTM, the default *encoder_randomaccess_vtm* configuration is chosen with intra-period=32 and GOP-size=32. To achieve a better compression ratio for the traditional codecs, we follow the test conditions in [Li *et al.*, 2023] to convert RGB videos to YUV444 to employ coding. The reconstructed YUV444 videos will be converted back to RGB for distortion calculation.

For LVC models, the intra-period and GOP sizes are set to the same value of 32. For uni-directional prediction LVC benchmark, DCVC-DC[Li *et al.*, 2023] and DCVC-FM [Li *et al.*, 2024] are selected. For bi-directional LVC, EEV-0.5 [Ye *et al.*, 2024] is evaluated. MPAI-EEV is a standard project focused on developing an AI-based End-to-End (E2E) video coding standard, aiming to enhance video compression by utilizing AI technologies, moving beyond traditional data coding methods. Since 2022, MPAI-EEV has released 5 verification models and a public benchmark for unmanned-aerial-vehicles (UAV) video coding [Jia *et al.*, 2023].

Experimental Results. For uni-directional predictive LVC models, including DCVC-DC and DCVC-FM, it can be observed that there is a noticeable performance gap compared to VTM for PSNR. In contrast, bi-directional predictive LVC models exhibit slightly lower PSNR performance than VTM but significantly outperform VTM for MS-SSIM, highlighting their superior perceptual quality. However, the performance of uni-directional predictive LVC models may be improved using larger GOP sizes. In low-latency coding scenario, uni-directional predictive LVC models demonstrate their strengths and achieve competitive performance.

6 Conclusion

This paper presents a comprehensive overview of recent progress in end-to-end optimized intelligent video coding models and their associated system-level developments. We systematically review both uni-directional and bi-directional prediction based LVC methods, detailing their architectural designs and optimization strategies. Additionally, we explore the utilization of LVC in practical systems using resource-constrained and energy-efficient platforms, further enhancing the versatility of LVC. We also present experimental results comparing its performance with traditional codecs, demonstrating significant improvements in compression efficiency. Through this analysis, we aim to illustrate the potential of LVC in the future of video coding and its applicability across diverse hardware environments.

Acknowledgments

References

[Alexandre *et al.*, 2023] David Alexandre, Hsueh-Ming Hang, and Wen-Hsiao Peng. Hierarchical b-frame video coding using two-layer canf without motion coding. In *CVPR*, 2023.

[Ballé *et al.*, 2018] Johannes Ballé, Nick Johnston, and David Minnen. Integer networks for data compression with latent-variable models. In *ICLR*, 2018.

[Boyce *et al.*, 2018] Jill Boyce, Karsten Suehring, Xiang Li, and Vadim Seregin. Jvet-j1010: Jvet common test conditions and software reference configurations, 2018.

[Bross *et al.*, 2021] Benjamin Bross, Ye-Kui Wang, Yan Ye, Shan Liu, Jianle Chen, Gary J Sullivan, and Jens-Rainer Ohm. Overview of the versatile video coding (vvc) standard and its applications. *IEEE TCSVT*, 2021.

[Chen *et al.*, 2017] Tong Chen, Haojie Liu, Qiu Shen, Tao Yue, Xun Cao, and Zhan Ma. Deepcoder: A deep neural network based video compression. In *VCIP*, 2017.

[Chen *et al.*, 2022] Meixu Chen, Todd Goodall, Anjul Patney, and Alan C Bovik. Learning to compress videos without computing motion. *SPIC*, 2022.

[Chen *et al.*, 2023a] Jiancong Chen, Meng Wang, Pingping Zhang, Shurun Wang, and Shiqi Wang. Sparse-to-dense: High efficiency rate control for end-to-end scale-adaptive video coding. *IEEE TCSVT*, 2023.

[Chen *et al.*, 2023b] Mu-Jung Chen, Yi-Hsin Chen, and Wen-Hsiao Peng. B-canf: Adaptive b-frame coding with conditional augmented normalizing flows. *IEEE TCSVT*, 2023.

[Cui *et al.*, 2021] Ze Cui, Jing Wang, Shangyin Gao, Tiansheng Guo, Yihui Feng, and Bo Bai. Asymmetric gained deep image compression with continuous rate adaptation. In *CVPR*, 2021.

[Dai *et al.*, 2022] Jincheng Dai, Sixian Wang, Kailin Tan, Zhongwei Si, Xiaoqi Qin, Kai Niu, and Ping Zhang. Nonlinear transform source-channel coding for semantic communications. *IEEE JSAC*, 2022.

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

[Fathima *et al.*, 2023] Noor Fathima, Jens Petersen, Guillaume Sautière, Auke Wiggers, and Reza Pourreza. A neural video codec with spatial rate-distortion control. In *WACV*, 2023.

[Feng *et al.*, 2020] Runsen Feng, Yaojun Wu, Zongyu Guo, Zhizheng Zhang, and Zhibo Chen. Learned video compression with feature-level residuals. In *CVPRW*, 2020.

[Feng *et al.*, 2021] Runsen Feng, Zongyu Guo, Zhizheng Zhang, and Zhibo Chen. Versatile learned video compression. *arXiv preprint: 2111.03386*, 2021.

[Ho *et al.*, 2022] Yung-Han Ho, Chih-Peng Chang, Peng-Yu Chen, Alessandro Gnutti, and Wen-Hsiao Peng. Canf-vc: Conditional augmented normalizing flows for video compression. In *ECCV*, 2022.

[Hong *et al.*, 2020] Weixin Hong, Tong Chen, Ming Lu, Shiliang Pu, and Zhan Ma. Efficient neural image decoding via fixed-point inference. *IEEE TCSVT*, 2020.

[Hu *et al.*, 2020] Zhihao Hu, Zhenghao Chen, Dong Xu, Guo Lu, Wanli Ouyang, and Shuhang Gu. Improving deep video compression by resolution-adaptive flow coding. In *ECCV*, 2020.

[Hu *et al.*, 2021] Zhihao Hu, Guo Lu, and Dong Xu. Fvc: A new framework towards deep video compression in feature space. In *CVPR*, 2021.

[Hu *et al.*, 2022] Zhihao Hu, Guo Lu, Jinyang Guo, Shan Liu, Wei Jiang, and Dong Xu. Coarse-to-fine deep video coding with hyperprior-guided mode prediction. In *CVPR*, 2022.

[Jia *et al.*, 2022] Chuanmin Jia, Xinyu Hang, et al. Fpx-nic: An fpga-accelerated 4k ultra-high-definition neural video coding system. *IEEE TCSVT*, 2022.

[Jia *et al.*, 2023] Chuanmin Jia, Feng Ye, Fanke Dong, Kai Lin, Leonardo Chiariglione, et al. Mpai-eev: Standardization efforts of artificial intelligence based end-to-end video coding. *IEEE TCSVT*, 2023.

[Kim *et al.*, 2023] Yeong-Chun Kim, Suyong Bahk, Seung Hwan Kim, Won Hee Lee, Dokwan Oh, and Hui-Yong Kim. Neural video compression with temporal layer-adaptive hierarchical b-frame coding. *arXiv preprint: 2308.15791*, 2023.

[Ladune *et al.*, 2020] Théo Ladune, Pierrick Philippe, Wassim Hamidouche, Lu Zhang, and Olivier Déforges. Optical flow and mode selection for learning-based video coding. In *MMSP*, 2020.

[Ladune *et al.*, 2021] Théo Ladune, Pierrick Philippe, Wassim Hamidouche, Lu Zhang, and Olivier Déforges. Conditional coding for flexible learned video compression. In *ICLRW*, 2021.

[Le *et al.*, 2022] Hoang Le, Liang Zhang, Amir Said, Guillaume Sautiere, Yang Yang, Pranav Shrestha, Fei Yin, Reza Pourreza, and Auke Wiggers. Mobilecodec: neural inter-frame video compression on mobile devices. In *ACM MMSys*, 2022.

[Li *et al.*, 2021] Jiahao Li, Bin Li, and Yan Lu. Deep contextual video compression. In *NeurIPS*, 2021.

[Li *et al.*, 2022] Jiahao Li, Bin Li, and Yan Lu. Hybrid spatial-temporal entropy modelling for neural video compression. ACM MM, 2022.

[Li *et al.*, 2023] Jiahao Li, Bin Li, and Yan Lu. Neural video compression with diverse contexts. In *CVPR*, 2023.

[Li *et al.*, 2024] Jiahao Li, Bin Li, and Yan Lu. Neural video compression with feature modulation. In *CVPR*, 2024.

[Lin *et al.*, 2020] Jianping Lin, Dong Liu, Houqiang Li, and Feng Wu. M-lvc: Multiple frames prediction for learned video compression. In *CVPR*, 2020.

[Lin *et al.*, 2021] Jianping Lin, Dong Liu, Jie Liang, Houqiang Li, and Feng Wu. Modulated variable-rate deep video compression. In *DCC*, 2021.

[Lin *et al.*, 2022] Chih-Hsuan Lin, Yi-Hsin Chen, and Wen-Hsiao Peng. Content-adaptive motion rate adaption for learned video compression. In *IEEE PCS*, 2022.

[Liu *et al.*, 2020] Dong Liu, Yue Li, Jianping Lin, Houqiang Li, and Feng Wu. Deep learning-based video coding: A review and a case study. *ACM CSUR*, 2020.

[Liu *et al.*, 2021] Haojie Liu, Ming Lu, Zhan Ma, Fan Wang, Zhihuang Xie, Xun Cao, and Yao Wang. Neural video coding using multiscale motion compensation and spatiotemporal context model. *IEEE TCSVT*, 2021.

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

[Lu *et al.*, 2020] Guo Lu, Chunlei Cai, Xiaoyun Zhang, et al. Content adaptive and error propagation aware deep video compression. In *ECCV*, 2020.

[Ma *et al.*, 2019] Siwei Ma, Xinfeng Zhang, Chuanmin Jia, Zhenghui Zhao, Shiqi Wang, and Shanshe Wang. Image and video compression with neural networks: A review. *IEEE TCSVT*, 2019.

[Ma *et al.*, 2022] Siwei Ma, Li Zhang, Shiqi Wang, Chuanmin Jia, Shanshe Wang, Tiejun Huang, Feng Wu, and Wen Gao. Evolution of avs video coding standards: twenty years of innovation and development. *Science China Information Sciences*, 2022.

[Ma *et al.*, 2023] Siwei Ma, Junlong Gao, Ruofan Wang, Jianhui Chang, Qi Mao, Zhimeng Huang, and Chuanmin Jia. Overview of intelligent video coding: from model-based to learning-based approaches. *Visual Intelligence*, 2023.

[Mao *et al.*, 2021] Hongzi Mao, Chenjie Gu, Miaosen Wang, Angie Chen, Nevena Lazic, Nir Levine, et al. Neural rate control for video encoding using imitation learning. In *ICMLW*, 2021.

[Mercat *et al.*, 2020] Alexandre Mercat, Marko Viitanen, and Jarno Vanne. Uvg dataset: 50/120fps 4k sequences for video codec analysis and development. In *ACM MMSys*, 2020.

[Pergament *et al.*, 2022] Evgenya Pergament, Pulkit Tandon, Oren Rippel, et al. Pim: Video coding using perceptual importance maps. *arXiv preprint: 2212.10674*, 2022.

[Pourreza and Cohen, 2021] Reza Pourreza and Taco Cohen. Extending neural p-frame codecs for b-frame coding. In *ICCV*, 2021.

[Rippel *et al.*, 2021] Oren Rippel, Alexander G Anderson, Kedar Tatwawadi, Sanjay Nair, Craig Lytle, and Lubomir Bourdev. Elf-vc: Efficient learned flexible-rate video coding. In *ICCV*, 2021.

[Sheng *et al.*, 2022] Xihua Sheng, Jiahao Li, Bin Li, Li Li, Dong Liu, and Yan Lu. Temporal context mining for learned video compression. *IEEE TMM*, 2022.

[Sheng *et al.*, 2024a] Xihua Sheng, Li Li, Dong Liu, and Houqiang Li. Spatial decomposition and temporal fusion based inter prediction for learned video compression. *IEEE TCSVT*, 2024.

[Sheng *et al.*, 2024b] Xihua Sheng, Li Li, Dong Liu, and Shiqi Wang. Bi-directional deep contextual video compression. *arXiv preprint arXiv:2408.08604*, 2024.

[Shi *et al.*, 2022] Yibo Shi, Yunying Ge, Jing Wang, and Jue Mao. Alphavc: High-performance and efficient learned video compression. In *ECCV*, 2022.

[Sun *et al.*, 2022] Heming Sun, Qingyang Yi, Fangzheng Lin, Lu Yu, Jiro Katto, and Masahiro Fujita. F-lic: Fpga-based learned image compression with a fine-grained pipeline. In *IEEE A-SSCC*, 2022.

[Van Rozendaal *et al.*, 2024] Ties Van Rozendaal, Tushar Singhal, Hoang Le, Guillaume Sautiere, Amir Said, Krishna Buska, et al. Mobilenvc: Real-time 1080p neural video compression on a mobile device. In *WACV*, 2024.

[Veerabadran *et al.*, 2020] Vijay Veerabadran, Reza Pourreza, et al. Adversarial distortion for learned video compression. In *CVPRW*, 2020.

[Wang *et al.*, 2016] Haiqiang Wang, Weihao Gan, Sudeng Hu, Joe Yuchieh Lin, Lina Jin, Longguang Song, Ping Wang, Ioannis Katsavounidis, Anne Aaron, and C-C Jay Kuo. Mcl-jcv: a jnd-based h. 264/avc video quality assessment dataset. In *IEEE ICIP*, 2016.

[Wu *et al.*, 2018] Chao-Yuan Wu, Nayan Singhal, and Philipp Krahenbuhl. Video compression through image interpolation. In *ECCV*, 2018.

[Xu *et al.*, 2023] Tongda Xu, Han Gao, Chenjian Gao, Yuanyuan Wang, Dailan He, Jinyong Pi, Jixiang Luo, Ziyu Zhu, Mao Ye, Hongwei Qin, et al. Bit allocation using optimization. In *ICML*, 2023.

[Yang *et al.*, 2020] Ren Yang, Fabian Mentzer, Luc Van Gool, and Radu Timofte. Learning for video compression with hierarchical quality and recurrent enhancement. In *CVPR*, 2020.

[Yang *et al.*, 2022] Ren Yang, Radu Timofte, and Luc Van Gool. Perceptual learned video compression with recurrent conditional gan. In *IJCAI*, 2022.

[Ye *et al.*, 2024] Feng Ye, Li Zhang, and Chuanmin Jia. Deep video compression with scaled hierarchical bi-directional motion model. In *ACM MM*, 2024.

[Yılmaz and Tekalp, 2021] M Akın Yılmaz and A Murat Tekalp. End-to-end rate-distortion optimized learned hierarchical bi-directional video compression. *IEEE TIP*, 2021.

[Zhang *et al.*, 2024a] Ping Zhang, Yiming Liu, Yile Song, and Jiaxiang Zhang. Advances and challenges in semantic communications: A systematic review. *National Science Open*, 2024.

[Zhang *et al.*, 2024b] Yiwei Zhang, Guo Lu, Yunuo Chen, Shen Wang, Yibo Shi, Jing Wang, and Li Song. Neural rate control for learned video compression. In *ICLR*, 2024.