Contents lists available at ScienceDirect

Cognitive Robotics

journal homepage: http://www.keaipublishing.com/en/journals/cognitive-robotics/

Recent trending on learning based video compression: A survey

Trinh Man Hoang, M.E[a], Jinjia Zhou, PhD[a,b,*]

[a] *Graduate School of Science and Engineering, Hosei University, Tokyo 1848584, Japan*
[b] *JST, PRESTO, Tokyo, Japan*

ARTICLE INFO

Keywords:
Video compression
Deep learning
Learning-based compression

ABSTRACT

The increase of video content and video resolution drive more exploration of video compression techniques recently. Meanwhile, learning-based video compression is receiving much attention over the past few years because of its content adaptivity and parallelable computation. Although several promising reports were introduced, there is no breakthrough work that can further go out of the research area. In this work, we provide an up-to-date overview of learning-based video compression research and its milestones. In particular, the research idea of recent works on learning-based modules for conventional codec adaption and the learning-based end-to-end video compression are reported along with their advantages and disadvantages. According to the review, compare to the current video compression standard like HEVC or VVC, from 3% to 12% BD-rate reduction have been achieved with integrated approaches while outperformed results on perceptual quality and structure similarity were reported for end-to-end approaches. Furthermore, the future research suggestion is provided based on the current obstacles. We conclude that, for a long-term benefit, the computation complexity is the major problem that needed to be solved, especially on the decoder-end. Whereas the rate-dependent and generative designs are optimistic to provide a more low-complex efficient learning-based codec.

Introduction

With the increasing of video content, nowadays, 70% of Internet traffic is used for video-based applications, including live streaming, low-latency realtime online communication, and video-on-demand platform. Meanwhile, video resolution and fidelity have also made huge steps forward (e.g., 4k, 8k, gigapixel [1], high dynamic range [2], bit-depth [3]) that make the dire situation worse. Therefore, many efforts have been invested in improving the video compression algorithms, whose task is to reduce the video size while keeping an acceptable visual quality reflected in the human visual system. With the benefit of transferring size reduction, video codec have been widely used in video-on-demand platforms like video streaming platforms (YouTube [4], Twitch [5]), online meeting systems (Zoom [6], Skype [7]) or online diagnosis systems [8,9] and so on [10,11] . Existing conventional video compression methods are well known for their hand-crafted artifacts. For instance, in the case of narrow bandwidth, the typical block-based video compression standards such as H.264/AVC [12], H.265/HEVC [13], and the incoming H.266/VVC [14] get involved in those artifacts by assigning large quantization parameter and coding unit size.

Recently, the development of neural networks (learning-based), especially convolutional neural networks (CNN) has risen the attention of researchers in the compression field. Different from conventional hand-crafted block-based coding, because learning-based approaches are possible to extract and utilize the features of the data, they could get better visual quality by avoiding blocktype artifacts. From the early period, image compression, which is the baseline of video compression, is firstly applied with learning-based

[*] Corresponding author.
E-mail addresses: trinhman.hoang.3f@stu.hosei.ac.jp (T.M. Hoang), zhou@hosei.ac.jp (J. Zhou).

https://doi.org/10.1016/j.cogr.2021.08.003
Received 6 July 2021; Received in revised form 16 August 2021; Accepted 16 August 2021
Available online 25 August 2021

methods. With the extraordinary performance of learning-based image compression compared to the conventional methods, from the past few years, learning-based video compression has been widely researched and got remarkable milestones. Hence, to provide a deep insight into current spots, trending directions, and the future development of learning-based video compression, this work presents a comprehensive review of video compression using neural networks.

There are many ways to apply the learning-based method to video compression. Start from the early beginning, they can be an integrated or replacement module of the conventional codec, then they further are an outside cooperated or guidance module. These modules have demonstrated their impressive compression performance and still been improving. Meanwhile, from the point, that researcher can somewhat overcome the non-differentiable quantization process in compression [15,16], learning-based end-to-end compression methods received a lot of scrutinies. In learning-based end-to-end compression, all components are learnable and linked to solving a global objective function, in compression the objective function is usually to represent the rate (number of bits) and distortion (quality) correlation. Different from the conventional codec, which is usually based on local optimum, the global objective function helps the learning-based end-to-end codec find the global optimized point, which, in theory, reveals a huge potential for further performance improvement and ondemand compression ability. Therefore, in this paper, we divide our survey into two main approaches:

- The conventional-learning-based cooperation approach. This approach contains the joint-processing of both conventional and learning-based methods, where the main component is usually the conventional codec. The learning-based methods can benefit the conventional codec as the inside module (intra-prediction, inter-prediction, in-loop filter), outside enhancement module (post-processing), or guidance (super-resolution-based, layered-based).
- The learning-based end-to-end compression approach. This approach contains only learnable components with a global objective function. However, their frameworks are very flexible and can be further separated into two major sub-approaches, predictive video coding, and generative video coding. The predictive video coding is usually designed to reflect conventional compression design based on the prediction and residual calculation. Whereas the generative video coding can be seen as an extension of the Variable Auto Encoder (VAE) technique.

The remainder of the article is structured as follows. In Section 2, we briefly summarize the learning-based compression via the image compression scheme since it can be seen as the basement of video compression. Section 3 covers the conventional-learning-based cooperation methods, including learningbased implementation on intra-prediction, inter-prediction, in-loop filtering, post-processing, and guidance modules. Section 4 provides a review of the learning-based end-to-end compression methods, including predictive video coding, generative video coding with the recent trending directions. We also discuss more their advantages, disadvantages, and brief comparison at the end of each section. Finally, Section 5 summarizes the current spot then points out the potential future research direction and concludes the paper.

Overview of the compression systems

Image compression is the primitive technique in the signal and the foundation of the video compression design.Meanwhile, video compression is likely an extended version of image compression with additional temporal relationship. Thus, in this section, we introduce a brief review of the conventional and learning-based image compression to provide you some background knowledge of the compression system design. Fundamentally, the conventional transformation image compression pipeline includes some basic modules, i.e., transformation, prediction, quantization, and entropy coding. The transformation was proposed for image compression to transform image pixel's value to compact and decorrelated coefficients. The transform function such as Fourier Transform [17], Hadamard Transform [18], or the commonly used Discrete Cosine Transform (DCT) [19] can condense the image energy in the low-frequency domain, therefore, led to better compression performance than directly use of image pixels value. Next, the prediction and quantization techniques are proposed to diminish the spatial and visual redundancy in images. In image compression, the prediction aims to provide the residuals of transformed components between neighboring blocks while video compression extends it by considering blocks from reference frames. Then, the quantization removes the least notable information by pruning less informative dimensions in the coefficient vector. Some methods have been introduced to improve the quantization like vector quantization [20] and trellis-coded quantization [21], but most of the existing image codecs round the quantization results to integers form to save the bit size for each value. After that, the decorrelated coefficients are compressed by entropy coding. Entropy coding will reduce the statistical redundancy within the processed vector, such as Huffman coding [22], arithmetic coding [23], Contentbased Adaptive Binary Arithmetic Coding (CABAC) [24]. For instance, the most common image standard, JPEG [25], uses DCT to transform each 8 × 8 partitioned image block. Then, a differential pulse code modulation (DPCM) [26] is applied to find the prediction residual between the DC components and their neighboring DCT blocks. A quantization matrix and rounding function is applied to the predicted residual. The quantized coefficients are then losslessly compressed by arithmetic entropy coding and sent to the channel.

Most of the existing block-based image and video codecs, such as JPEG, JPEG2000 [27], BPG [28] for the image compression and AVC/H.264, HEVC/H.265 or the incoming VVC/H.266 for the video compression, are blockdependent and each block is sequential-dependent also. Those attributes are the cause of the block artifacts [29–31] and prevent the codecs from the parallel computational platform. In addition, each module is usually independently designed and optimized, which also limits the compression performance because of the local optimization. Meanwhile, with the benefit of GPU development for parallel computing, the neural network has achieved great improvement in many fields. Especially, convolutional neural network and their extended designs have been demonstrated that can well learn the image spatial features that can serve to many image and video processing tasks [32–35,37].

However, it is not trivial to add the CNN model into the compression framework. Training a CNN requires the differentiability of the loss function concerning trainable parameters of the CNN. Nonetheless, the quantization module of the lossy compression framework results in zero gradients almost everywhere that terminate the parameters updating in the CNN training process. Early works in learning-based image compression let the conventional codec do that quantization part on the predicted residual between the learning-based reconstructed image and the original image. The lossless latent from the bottleneck layer of the autoencoder network and the quantized residual then were used to reconstruct the image on the decoder [36–39]. Inspired by the success of CNN-based super-resolution, [40–42] further improve the framework by firstly down-sampling the original image to a smaller scale then feeding it into the conventional codec. A post-processing model is added to the output of the conventional codec to fulfill the loss of information. However, since the image is processed over two lossy modules, the residual is extremely worse without any specific training process. By pointing out that the main informative information should not be lost and an image can be divided into different aspects, in recent years, layered image compression has achieved outstanding achievement [43–46]. In layered image compression, an image is usually separated or transformed into two components: the most informative part, the reconstruction-aid information. The most informative part is usually dense and compact and will be losslessly compressed to preserve important information. Meanwhile, the reconstruction-aid information and the residual can be lossy compressed based on the bit-rate requirement. The reconstruction-aid information is varied from the color-mask [47], semantic information [43,46,48,49], to attention mask [50–53]. The most common reconstruction-aid information is the semantic mask which is outperforming the conventional image codec like JPEG or BPG. This approach is still catching the researcher's eyes that [46] recently has demonstrated that some reconstruction-aid information does not need to be sent on the channels which can save a sufficient amount of bitrates.

According to the BD-rate reduction [54], which represents the ratio of transferring bits reduction at the same reconstructed quality, the above approach has got noticed results in compression ratio according to the BD-rate. However, their drawbacks are obviously seen, the additional computation of learningbased models plus the sequential processing with the conventional codec leads to the extremely heavy codec for both encoder and decoder sides. Even worse, the learning-based module usually falls into the local optimal for distortion quality with limited bit-rate optimization since it depends on a non-trainable conventional codec. Addressed these problems, researchers are giving more attention to end-to-end learning-based image compression recently. Given an input image x with its distribution $\mathrm{p_x}$, the encoder with the transformer E and the quantizer Q, the discrete quantized code y and the reconstructed x are then generated as following:

$$\hat{y} = Q\left(E\left(x, \theta_e\right)\right) \quad (1)$$

$$\hat{x} = IE\left(Q\left(E\left(x, \theta_e\right)\right), \theta_{ie}\right) \quad (2)$$

Where IE is the invert transform on the decoder and θ_e, θ_{ie} denote the parameters of E and IE, respectively. The end-to-end compression takes both distortion D and compression bitrate R into the loss function as follow:

$$L = R(\hat{y}) + \lambda D(x, \hat{x}) \quad (3)$$

Where D can be any distortion function(MSE [55], SSIM [56], MS-SSIM [1], perceptual function [15], etc.), λ is the hyper-parameter indicates the ratedistortion trade-off, and R can be formulated as the cross-entropy from an estimation entropy model. Since Ball´e et al. [15,57], introduced learnable analysis and synthesis transformation network GDN along with the uniform noise quantization, the attention on learning-based end-to-end image compression has increased rapidly. Ball´e et al. [58] applied the recurrent models for variablerate learned compression or [59] with enhanced nonlinearity. Other advanced CNN architectures [36,51,60–64] have also improved the transformation capacity. Meanwhile, differentiable quantization received a lot of notices, Cai and Zhang [16] proposed soft-to-fine scalar vector quantization or [65] proposed a model to learn the quantization parameters. Nonetheless, uniform noise adaption from [15] is still the most used training-quantizer in recent works due to its convenience and efficiency. After that, the quantized latent is further compressed by the entropy coding, which requires an entropy model to estimate the probability distribution, some milestones are [66,67] with the hyperpriors, [63,68–71] with predictive models, or [59,72,73] with learned parametric models. Specifically, hyperpriors-based entropy estimator has got a lot of improvement and adaption in recent works [74–76]. The reader may also refer to [77] for more details and research progress on the learning-based endto-end image compression. In this work, we briefly introduce it to provide the baseline knowledge of the learning-based video compression.

Conventional-learning based cooperation

Like image compression, research in learning-based video compression also starts from the cooperation between the conventional codec and the learningbased modules. In this section, we briefly introduce the progress of applying learning-based techniques into the conventional codec major components like intra prediction, inter prediction, in-loop filtering, and additional learnable guiding module of the codec. Even the in-loop filtering and postprocessing modules are different by the position corresponding to the codec, their task is closely similar to each other, so we review them in the same sub-section for a comprehensive presentation.

Integrated deep tools

Intra prediction

Similar to images, video frame-block is spatially correlated with the neighboring samples. Therefore, based on the current block, the next neighboring blocks can be predicted and only the predicted residuals (errors) are sent. Existing conventional codecs provide

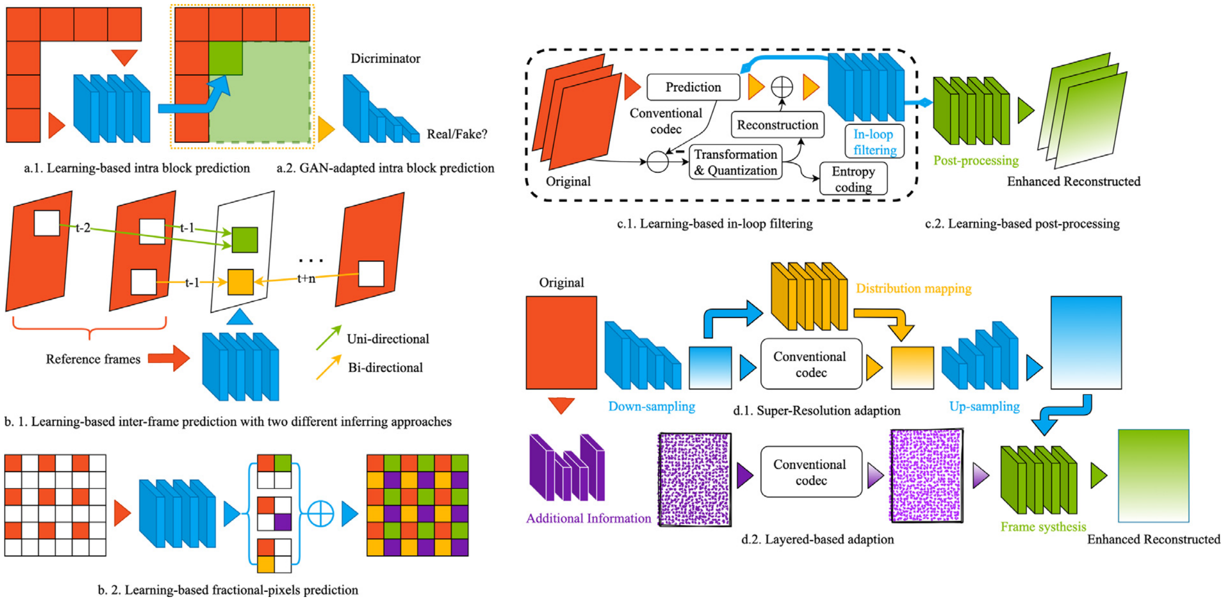

Dicriminator
Real/Fake?
a.1. Learning-based intra block prediction
a.2. GAN-adapted intra block prediction
Prediction
Conventional codec
Reconstruction
In-loop filtering
Transformation & Quantization
Entropy coding
Original
Post-processing
Enhanced Reconstructed
c.1. Learning-based in-loop filtering
c.2. Learning-based post-processing
t-2
t-1
t-1
t+n
Reference frames
Uni-directional
Bi-directional
b. 1. Learning-based inter-frame prediction with two different inferring approaches
Original
Distribution mapping
Conventional codec
Down-sampling
d.1. Super-Resolution adaption
Up-sampling
Additional Information
Conventional codec
d.2. Layered-based adaption
Frame sythesis
Enhanced Reconstructed
b. 2. Learning-based fractional-pixels prediction

Fig. 1. a.1. Basic learning-based block intra prediction and its GAN-based adaption a.2. with a discriminator for the whole predicted frame; b.1. Basic learning-based frame inter prediction with uni-direction and bi-direction approaches; b.2. The basic concept of learningbased fractional-pixels prediction; c.1. The basic concept of learning-based in-loop filtering; c.2. The post-processing processes the decompressed frame outside of conventional codec; d.1. Basic super-resolution adaption on video coding and its layered-based extension d.2

several intra prediction modes according to the applied directional functions on the surrounding samples. In practice, the predefined mode with the best prediction in term of R-D optimization is selected.

Naturally, we expect that a better prediction will lead to a better coding performance. Hence, several CNN-models have been applied as the block predictor for the conventional codec [78–80] (see Fig. 1.a.1), other neural network models liked GAN [81,82] (see Fig. 1.a.2) or RNN [83–85] have also been implemented with the same intuition that is a better predicted quality will reduce the residual error. The major difference comes from the input data, most of the works leverage the current block and neighboring blocks, others used the whole frame in different input directions [83,85], the predicted results from other standard modes have also been considered as inputs for the networks [80,86]. Compare to HM [87] software, the reduction on BD-rate [54] of recent CNN-based [80], RNN-based [85] and GAN-based [82] models are 3.4%, 2.65% and 1.2% respectively on Y-channel over HEVC codec with QP = 22, 27, 32, 37 with a trade-off of reported extremely high in complexity. Meanwhile, the whole frame is also used to predict the best standard modes [88–90]. By directly chose the prediction mode according to the output of the model, more than 60% intra-coding time of anchor codecs can be saved with a marginal increase of BD-rate [90].

Inter prediction

Temporal information is the key factor that distinguishes the video from the image. In typical hybrid video compression, the temporal information is defined as the motion estimation on the references coded frames against the current frame. The reference blocks can come from past (forward/uni-prediction) or both past and future (bi-directional prediction). In a typical conventional codec, the best motion vector (MV), which results in fewer residuals left on the matching block, is selected. The precision of MV can be further improved by leveraging the higher fidelity and more fine-grained motion compensation, e.g., fractional-pixel interpolation, discrete cosine transform-based interpolation filter (DCTIF) [91].

The early research direction for learning-based inter-prediction is using CNN to capture the references blocks features. Those features are then used to enhance the predicted block, therefore reduce the residual errors (see Fig. 1.b.1). Because of the reference block's appearance, both temporal and spatial features are used for the enhancement [92–94] and can reduce up to 2.9% BD-rate [94] on the Y channel compared to the HM Random Acess configuration. For the HEVC specific adaption, CNN has been implemented for the factional-pixel interpolation [95–98] (see Fig. 1.b.2), however, their ability of generalization is very dependent, i.e., the FRCNN [95] need 120 models for 4 common quantization parameters (QP). Meanwhile, [99–102] proposed using CNN to generate a virtual reference frame that closer to the current frame whose inputs come from the coded frames. In addition, instead of using the traditional simple average, [103,104] used CNN on the sub-pixels to infer the inter-prediction block in a non-linear approach.

In-loop filter and post-processing

In-loop filtering (ILF) in cooperated with Sample Adaptive Offset (SAO) modules [105–110] are introduced to the video coding standard to reduce the block artifacts caused by the block-based processing manner of the conventional codec. The ILF and SAO,

Table 1
Results on BD-rate reduction (%) from recent works of the conventional-learning based cooperation approach in comparison with their anchors.

Module	Published	Refs.	BD-rate(%)	Anchor
Intra	2019	[80]	-3.4	HM 16.9 AI
prediction	2019	[85]	-2.65	HM 16.15 AI
	2021	[81]	-1.2	HM 16.15 AI
Inter	2018	[104]	-3	HM 16.15 RA
prediction	2019	[98]	-1.2	HM 16.7 RA
	2020	[94]	-2.9	HM 16.6 RA
Inloop	2021	[121]	-5.06	VTM 9.3 RA
Filtering	2021	[124]	-4.03	VTM 6.0 AI
Post	2019	[133]	-9.76	HM 16.0 LDP
processing	2019	[140]	-6.16	HM 20.0 LDP
	2020	[130]	-6.7	VTM 7.0 RA
Guidance	2020	[153]	-6.2	HM 16.20 RA
	2021	[149]	-12.6	VTM 4.01 RA

therefore, improve the quality of the predicted frame that leads to a smaller residual error and bitrate. Since the process can be seen as an enhancement problem, many CNN models have been transferred to this task. The early works mostly depend on spatial information [111–116], later on, temporal information steps into the inputs by using the references frames [117–120]. Most works are the replacement of the conventional module (see Fig. 1.c.1), others can be added in the middle of ILF and SAO or after both. The most recent CNN-based ILF/SAO can achieve up to 12.6% [119] and 5.06% [121] BD-rate reduction compared to HM and VTM [122] Random Access configuration, respectively. Similar to other learning-based modules, CNN-based ILF/SAO is very dependent on the compression configuration, therefore it requires a lot of memory to store huge numbers of model weights and increase the coding complexity enormously. Later work [123] has tried to combine all QPs in their training phase to use only one model with a marginal loss on performance was reported, or [124] with a QP-attention module to detect different compressed noise levels.

Because ILF and SAO are integrated inside of the encoder and decoder, existing compressed video cannot be benefited from the new technique. Hence, some post-processing methods have been proposed to perform only on the decoderend (see Fig. 1.c.2). At first, several works with different CNN-based models [125–131] use supervised learning to perform the enhancement based on spatial information. Next, some priors, which are the cause of artifacts, have been extracted from the sending bitstream and inputted to the CNN [132–134]. Then, the multi-frame enhancement approach uses the higher quality frames to enhance the lower one [135–138]. The performance is still increasing due to the complicated network architectures, up to 0.96dB of PSNR [56] increase compare to HM codec recently [139] with the multi-frames approach. Meanwhile, several works tried to leverage more information while keeping a limited number of parameters [139–141].

Learning-based guidance of the conventional codec

The conventional codec usually compressed a video in a uniform process. To specify the compressed process adaptively, learning-based methods have been implemented in pre-processing/integrated in-cooperate with the post-processing side. With the very low-bitrate compression, super-resolution is the common approach [142–146]. Generally, the down-sampled model on the pre-processing or integrated positions compacts the input frame into the condensed and informative scaled version. Then, after the conventional codec output the compact reconstructed frame, the up-sampled and restoration models will fulfill the down-sampled loss (see Fig. 1.d.1). More temporal information and be leveraged to enhance the restoration based on the motion compensation [147,148] or scaling order [149–152]. Later on, GAN-based methods [153–155] also implemented into the host codec with the similar position of super-resolution methods but to address the perceptual quality instead of frame distortion. For the Region-Of-Interests-based (ROI) compression, the attention mask [156,157], semantic mask [158–160], foreground/background mask [161,162] can be separated or embedded compressed with the down-sampled frame, i.e., layered-based compression (see Fig. 1.d.2). Those priors are used to reduce the unimportant part of the frame or finding a better MV at the multi-frame enhancement phase.

Discussion

Overall, the improvement of compressing performance is feasible when applying learning-based modules in cooperated with conventional codecs. Table 1 shows the quantitative results and Table 2 briefly presents the proposals of some remarkable recent works in this conventional-learning based cooperation approach. We can see that most basic conventional module tasks can be handled by learning-based models like intra- prediction, inter-prediction, in-loop filtering, or even post-processing for existing compressed video. Learning-based modules can further improve the demand-flexibility of the conventional codec by adaptive pre-, integrated, and post- correlated processing. The replaceable ability and demand-flexibility allow researchers to transfer them to the newest video coding standards, i.e., H.266/VVC, AVS3 [163].

Nonetheless, the compression performance of learning-based modules comes with a huge trade-off of complexity and storage memory increment. For example, [100] virtual frame interpolation can achieve 4.6% of BD-rate reduction while increasing 35% and

Table 2
Research idea from recent works of conventional-learning-based cooperation approach.

Module	Published	Refs.	Method
Intra	2019	[80]	Enhance the anchor's intra-prediction result using multi-scale CNN
Prediction	2019	[85]	Use Vertical and Horizontal RNN extract neighboring feature for the prediction
	2021	[81]	Use GAN-based technique to generate predicted block from neiboring blocks
Inter	2018	[104]	Use CNN on the sub-pixels to infer the inter-prediction block
Prediction	2019	[98]	Learning-based fractional interpolation
	2020	[94]	Generate a Virtual Reference Frame as the closest neighbor of the current frame.
Inloop	2021	[121]	Enhance the predicted frame using a specific multi-density CNN
Filtering	2021	[124]	Perform the In-loop filtering with only one training weight set by levering a QP attention module
Post	2019	[133]	Leverage the Block-information for the enhancement
Processing	2019	[140]	Reduce the network complexity when leveraging more information
	2020	[130]	Supervise learning with specificMFRNet architecture
Guidance	2020	[153]	GAN and Super-resolution based adaption for the video bit depth reduction
	2021	[149]	Super-resolution based coding with anchor's QP dependent

4276% of encoding and decoding time of the HM codec, respectively. That is the main reason that there are almost no learningbased modules as the main part in any incoming video compression standards like VVC or AVS3. Since more prior information has been inputted into the network and the video resolution rise, the complexity problem becomes much more serious [164]. Some works have addressed the complexity problem by designing a lightweight network [140,165–168], others reduced the storage by adaptive online-training manner [169,170] or using only one generalized trained model [171,172]. However, none of them achieve the real-world application required complexity or without performance decrease, therefore more study in this direction is in need.

End-to-end learning-based video compression

Although learning-based modules can enhance the conventional codec performance, they also inherit the local optimization problem from the conventional modulo design. Thanks to the development of the differentiable quantization function [15,57,16,65], recently, the learning-based end-to-end (ETE) video coding receives a lot of studies. Like conventional video codec, learning-based ETE codec compresses the intra-frame with the learning-based ETE image compression methods (see Section 2). The most different comes from the inter-frame processing methods. Currently, there are two primitive approaches to infer the inter-frame: Predictive and generative video coding. In this section, we introduce the key idea of each approach, its pros/cons, and the current status. More discussion about the trending research direction and other side improvements are also presented.

Predictive Video Coding

Inspired by the conventional predictive design, this approach's general framework is an imitator that fully replaces the conventionally designed modules with learning-based models but mostly processes on the whole video frame instead of a divided block. Similar to the conventional codec, the key idea is to compress the residual errors between the predicted frame and the current frame. In specific, while intra-frame leverages the learning-based ETE image compression, the motion estimation of inter-frame prediction is replaced with the learningbased flow-estimation [174–178], the motion compensation become warping function [176,179–181], and the learning-based frame synthesis model will do the reconstruction task [182–184]. The estimated flow and the residual error then are compressed and sent to the decoder. Especially, with the join of the differentiable quantization methods, all modules are linked together to perform a global optimization training process for the R-D loss (see Eq. (3)). Using warping function W, given the decompressed flow $\hat{f}$ and residual $\hat{r}$, the reconstructed form of frame at decoded time t is as following:

$$\hat{x}_t = W\left(\widehat{x_{t-1}}, \hat{f}_t\right) + \hat{r}_t \tag{4}$$

The most well-known framework in this approach recently is DVC [173] (see Fig. 2), which is the first that successfully replaces all conventional modules with learning-based models. Based on its framework, more improvement has been introduced. [185–187] increased the number of reference frames, [188,189] performed the motion and residual prediction on feature level, [92,190] pre-defined the decoded order to perform the hierarchical quality coding. Then, more conventional coding idea has also been implemented, from MV difference compression [185], block-level processing [191], to in-loop interpolation [192]. Besides, the sequences model is widely used to leverage the long-term memory [190] or feedback recurrent approach [187,191] for the current frame prediction. Later, researchers are taking more attention to network architecture design purpose to adapt to the compression scheme instead of the vanilla model transferring [176,193,194]. The most noticed work recently is the scale-space-flow estimation [176], that successfully produced a better MV with different Gaussian blurring filters for the compression task by considering residual errors instead of the accuracy of flow pixels value under the supervised pre-trained weight. Using a fixed-resolution scale-space volume $X = [x, x * G(\sigma_0), x * G(2\sigma_0), \ldots, x * G(2^{M-1}\sigma_0)]$, where $x * G(\sigma)$ indicates the convolution of x with a Gaussian kernel σ, each 3-channel field of X is then defined as $g := (g_x, g_y, g_z)$, the prediction is as following process:

$$\begin{gathered} \hat{x} = \mathrm{Scale} - \mathrm{Space} - \mathrm{Warp}(x, g)\ \mathrm{s.t} \\ \hat{x}[x, y] = X[x + gx[x, y], y + gy[x, y], gz[x, y]] \end{gathered} \tag{5}$$

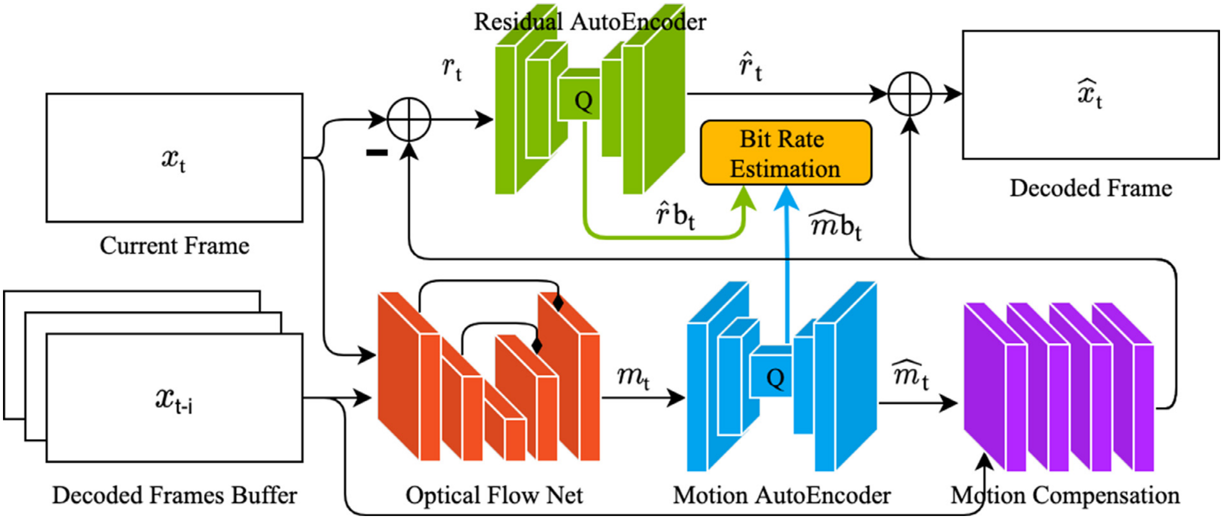
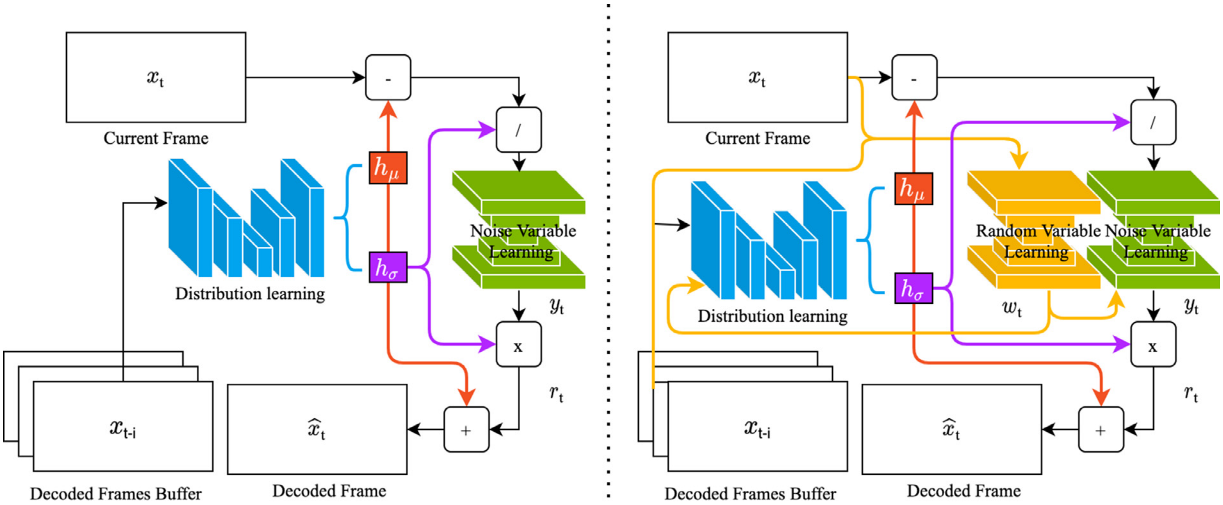


Fig. 2. DVC [173] end-to-end predictive coding framework. m, r represent the motion flow and estimated residual with their corresponding quantized bitstream m̂b, r̂b and their reconstructed version m̂, r̂, respectively.



a. Basic generative video coding

b. A hierarchical adaption for generative video coding

Fig. 3. a. Basic generative video coding framework. b. An adaptive generative video coding for a random variable w [195].

Checking their visual results [176], it is obvious that the framework has chosen the high Gaussian blurring scale for the large motion area instead of trying to predict an unpredictable flow. Therefore, the residual error is intuitively smaller than using an incorrectly predicted frame.

Generative video coding

Different from the predictive approach where the major task of the framework is to produce the best-predicted frame with the lowest residual error, the primitive task of generative video coding framework is to learn the parameter of probability distribution p_y for the latent representation. Fig. 3.a shows the pure version of the video coding with only temporal autoregressive transform. With further extend with a random variable, the general reconstruction form of a generative video coding framework is as following:

$$\hat{x}_t = h_\mu\left(\widehat{x_{t-1}}, w_t\right) + h_\sigma\left(\widehat{x_{t-1}}, w_t\right) \odot g_v\left(v_t, w_t\right) \tag{6}$$

Where h_μ and h_σ are functions that transform the $\widehat{x_{t-1}}$ and sub-latent w_t into mean and variance parameters, respectively. The function $g_v(v_t)$ convert the received latent $[w_t, v_t]$ into a noise variable the represent the residuals reflect to the $h_\mu(\widehat{x_{t-1}}, w_t)$ predictive output. It is worth to mentioned that even generative coding also adapts the prediction process, the learning information is obviously different from the native predictive coding.

Since the $\mathrm{p_y}$ is unknown, a learnable $\mathrm{q_y}$ is used as an estimation of $\mathrm{p_y}$. Therefore, the rate term R now becomes the cross-entropy of $\mathrm{p_y}$ and $\mathrm{q_y}$:

$$R = E\left[-log_{qy}(y)\right] \tag{7}$$

Where E is the entropy coding function [22–24]. The $\mathrm{q_y}$ can be learned by the parameter of the variational autoencoder network [196–198] directly from the training frames and autoregressive is the common inferring way with a fixed [195,198] or a dynamic number of priors [196,197]. Inherited from image compression, using GDN transform [15,57], the latent distribution $\mathrm{p_y}$ in video compression is also commonly pre-defined to obey the Gaussian distribution [199] with a learnable mean and standard deviation. These parameters are then adaptive predicted and become the prior information to the variational autoencoder [66]. Many recent works are studying to improve this module by a more complex design [74,200], a context-adaptive design [71] or including the temporal information via sequences-learning technique [201,202]. Guo et al. [195] recently explore the ability of the hyperprior on the random variables that can have an effect on the residual error rate of the predictive coding approach, the results have demonstrated very promising potential for a more complex combination of two approaches. By replacing the $\mathrm{h_\mu}$ in equal 6 with W in equal 4, an generative flow-based prediction coding was conducted. This is feasible since both function are learnable and output a prediction result. Furthermore, w is also used as a hyper latent to further improve learned priors for residual v entropy model (see Fig. 3.b). Meanwhile, GAN-based training have also widely used [203,204] for the perceptual quality instead of the quantitative distortion on the extremely low bitrate compression.

Discussion

While the conventional-learning-based framework is usually strong on the mean square error (MSE) or Peak Signal Over Noise Rate PSNR because of its conventional baseline, the learning-based ETE framework is surpassed on the structure similarity by avoiding the block artifacts since they process with fully spatial information of the frame. The early works of both ETE approaches [141,173,196] have achieved a comparable or even better performance than the H.265/HEVC with the default setting on SSIM [56] or MS-SSIM [205] evaluation metric. The distance on SSIM/MS-SSIM becomes more distinct when more prior is fed into the framework [185,201], especially on smooth and mid-motion sequences [206]. In comparison on the PSNR evaluation metric, recent works [190,195,197,201] got the comparable or marginal increase compared to the H.265/HEVC with low delay P configuration under the fast preset on several common test sets [206–209]. Other than the PSNR for quantitative distortion and SSIM for the structure similarity, LPIPS [210], FID [211], and KID [212] are usually used to evaluate the perceptual quality of the GAN-based framework. Visually, GAN-based results [203,204] which have high perceptual quality is in a better quality even have a lower score on PSNR and MSSIM, especially for the extreme low bitrate.

So far, the end-to-end learning-based video compression methods have surpassed traditional codecs on MS-SSIM, perceptual quality, or even PSNR on low bitrate H.265/HEVC codec. However, no work can sufficiently beat the conventional codec on high bitrate compression, especially the incoming H.266/VVC. Although VVC is known for a much more computation requirement, by the design approach, only its encoder suffers that burden, which is suitable for the modern server-client application. Meanwhile, the learning-based ETE compression requires powerful hardware i.e, GPU, to process the compression on both the encoder and decoder sides. The problem becomes more serious since more prior is being leveraged recently required a much more complex network design. Therefore, researchers now should take the real application scenario into the framework and network design, from the end-user view, lightweight network architecture and unbalance framework are needed to carefully study. Several works have been exploring this problem, mainly using the model compression methods [213].

Summary, conclusion and future note

This work presents an overview of the updated methods of learning-based video compression research. From the survey produced earlier in this paper, the conventional-learning-based modules have successfully been implemented and can improve the compression rate in most of the existing conventional codecs, including H.264/AVC, H.265/HEVC, and H.266/VVC. Whereas the learningbased end-to-end approach has achieved comparable distortion efficiency on PSNR while outperforming on MS-SSIM and perceptual quality compares to H.265/HEVC under some specific settings. However, there is no learning-based end-to-end method that can reach the VVC performance on the distortion evaluation metric or even HEVC performance on high bitrate compression. Based on our survey, the advantages of learning-based video compression are mainly four-folds. First, since the learning-based model is content-adaptive to the huge amount of training data, it easy to surpass the handcrafted designed module on specific tasks. Second, the difference from the conventional codec, the learningbased models usually explore the large receptive field in both spatial and temporal domains, therefore provides a more accurate prediction or latent distribution. This manner also helps the codec to avoid the blocking artifact and become flexible in temporal exploration. Third, the direct linkable ability allows the learning-based modules to perform the global optimization that is the potential factor for further improvement on the R-D trade-off and specific human vision task. Finally, the flexibility of the learning-based method allows them to quickly inherit the newest technology, extend the design and transfer the knowledge easily.

In the trade-off for the compression ratio performance, current learningbased video compression methods are facing many obstacles that are required to be further investigated:

- **Complexity and memory requirement.** One of the major limitations of the learning-based approach compared to the conventional one is the enormous burden of computation and memory requirements. The current learning-based model requires too much computational power to achieve a marginal increase of compression performance [100,164]. It is even worse since recent works [185,195,201] try to increase the compressed performance by leverage more priors that require much more complex and deeper network architecture. Different from the conventional codecs which keep the decoder complexity be light on CPU processing, the learning-based method usually introduces more heavy computation for both encoder and decoder, which requires powerful hardware, i.e., GPU on the end-user side [77]. Meanwhile, the memory required to store all the trained weights is also a big problem because of the varied compression schemes. Currently, several works have addressed this problem by designing a lightweight network [140, 165,166,168] in co-operation with an online training scheme [169,170] or using model compression methods for an unbalance decoder [167]. For memory saving, one-model-only is the most common research direction [123,124,172]. However, no work can beat the incoming VVC codec in both compressed performance and complexity on the decoderend, especially on CPU processing. With the rapid increase of video resolution (4K, 8K) and varied video-based demand applications, the learningbased design needs a more flexible and lightweight design, especially on the decoder-end side.

- **Rate-dependent model design.** The most common framework design of current learning-based works intuitionally assumes that distortion of the predicted frame and rate are in a linear correlation. Hence, the network is designed to produce a better-quality predicted frame that hopefully reduces the compression rate of the residual. In fact, a quantitative related small residual error does not linearly lead to a small entropy compressed bitstream which was recently demonstrated by [176,193,194]. By taking the entropy compression algorithm into account of network design and training strategy, it is potential to have an efficient framework with a lightweight prediction network that needs only to produce a residualfriendly predicted frame instead of a very accurate one.
- **The generative design for the learning-based model.** So far, the generative approach has demonstrated its performance on many computer vision tasks. Its advantage becomes clearer in the compression field where the learned distribution can save a lot of sending information. Especially in the video scheme, the generative model can leverage the autoregressive processing of the temporal information to get a much more accurate learned probability distribution recently [195,201]. With the discovery from [195] that generative design can be implemented for any random variable concerning the residual error, the generative model is now waiting to be adapted into the incoming studies.

In conclusion, learning-based video compression has achieved a noticeable improvement over the past few years. Learning-based modules have been widely embedded into the conventional codecs to further improve their prediction performance or enhance their results. Meanwhile, learning-based end-to-end video compression has received many studies and can get a comparable compression performance with conventional codec on distortion metric or better on perceptual or structure quality. Nonetheless, because of their huge burden on computation, especially on the decoder-end, none of them appear in the industry application so far. Therefore, complexity reduction is one of the most crucial research directions, along with rate-dependent design direction and generative design, which are to find a more effective low-cost framework for the R-D optimization.

References

[1] D.J. Brady, M.E. Gehm, R.A. Stack, D.L. Marks, D.S. Kittle, D.R. Golish, E. Vera, S.D. Feller, Multiscale gigapixel photography, Nature 486 (7403) (2012) 386–389.

[2] S.B. Kang, M. Uyttendaele, S. Winder, R. Szeliski, High dynamic range video, ACM Trans. Graph. (TOG) 22 (3) (2003) 319–325.

[3] M. Winken, D. Marpe, H. Schwarz, T. Wiegand, Bit-depth scalable video coding, in: Proceedings of the IEEE International Conference on Image Processing, Vol. 1, IEEE, 2007, pp. I–5.

[4] C. H. Steve Chen, Youtube, https://www.youtube.com, 2005, accessed 24 June 2021.

[5] E. S. Justin Kan, Twitch, https://www.twitch.tv, [Online; accessed 24June-2021] (2021).

[6] Z. V. C. Eric Yuan, Zoom, https://zoom.us, [Online; accessed 24-June2021] (2021).

[7] S. T. M. Priit Kasesalu, J. Tallinn, Skype, https://www.skype.com/en/, [Online; accessed 24-June-2021] (2021).

[8] Mission Health, https://missionhealth.org/mission-telemedicine/, [Online; accessed 24-June-2021] (2021).

[9] Apollo 247, https://www.apollo247.com/specialties, [Online; accessed 24-June-2021] (2021).

[10] Insecam, http://www.insecam.org, [Online; accessed 24-June-2021] (2021).

[11] SAMSUNG, Samsung's family hub, https://www.samsung.com/us/explore/family-hub-refrigerator/overview/, 2021, accessed 24 June 2021.

[12] T. Wiegand, G.J. Sullivan, G. Bjontegaard, A. Luthra, Overview of the h. 264/avc video coding standard, IEEE Trans. Circuits Syst. Video Technol. 13 (7) (2003) 560–576.

[13] G.J. Sullivan, J.R. Ohm, W.J. Han, T. Wiegand, Overview of the high efficiency video coding (hevc) standard, IEEE Trans. Circuits Syst. Video Technol. 22 (12) (2012) 1649–1668.

[14] S. K. J. Chen, Y. Ye, Algorithm description for Versatile Video Coding and Test Model 11, https://jvet-experts.org/doc_end_user/current_document.php?id=10541/, [Online; accessed 24-June-2021] (2021).

[15] J. Ball´e, V. Laparra, E.P. Simoncelli, End-to-end optimization of nonlinear transform codes for perceptual quality, in: Proceedings of the Picture Coding Symposium (PCS), IEEE, 2016, pp. 1–5.

[16] J. Cai, L. Zhang, Deep image compression with iterative non-uniform quantization, in: Proceedings of the 25th IEEE International Conference on Image Processing (ICIP), IEEE, 2018, pp. 451–455.

[17] E. Agustsson, F. Mentzer, M. Tschannen, L. Cavigelli, R. Timofte, L. Benini, L. Van Gool, Soft-to-hard vector quantization for end-to-end learning compressible representations, Adv. Neural Inf. Process. Syst. 30 (2017) 30 NIPS.

[18] R.N. Bracewell, R.N. Bracewell, The Fourier Transform And its Applications, Vol. 31999, McGraw-Hill, New York, 1986.

[19] W.K. Pratt, J. Kane, H.C. Andrews, Hadamard transform image coding, Proc. IEEE 57 (1) (1969) 58–68.

[20] N. Ahmed, T. Natarajan, K.R. Rao, Discrete cosine transform, IEEE Trans. Comput. 100 (1) (1974) 90–93.

[21] R. Gray, Vector quantization, IEEE ASSP Mag. 1 (2) (1984) 4–29.

[22] T.R. Fischer, M.W. Marcellin, M. Wang, Trellis-coded vector quantization, IEEE Trans. Inf. Theory 37 (6) (1991) 1551–1566.

[23] D.E. Knuth, Dynamic huffman coding, J. Algorithms 6 (2) (1985) 163–180.

[24] I.H. Witten, R.M. Neal, J.G. Cleary, Arithmetic coding for data compression, Commun. ACM 30 (6) (1987) 520–540.

[25] D. Marpe, H. Schwarz, T. Wiegand, Context-based adaptive binary arithmetic coding in the h. 264/avc video compression standard, IEEE Trans. Circuits Syst. Video Technol. 13 (7) (2003) 620–636.

[26] W.B. Pennebaker, J.L. Mitchell, JPEG: Still image data compression standard, Springer Science & Business Media, 1992.

[27] C. Harrison, Experiments with linear prediction in television, Bell Syst. Tech. J. 31 (4) (1952) 764–783.
[28] M. Rabbani, Jpeg2000: Image compression fundamentals, standards and practice, J. Electron. Imaging 11 (2) (2002) 286.
[29] F. Bellard, BPG Image format, https://bellard.org/bpg/, 2018, accessed 24 June 2021.
[30] T. Vlachos, Detection of blocking artifacts in compressed video, Electron. Lett. 36 (13) (2000) 1106–1108.
[31] S. Ye, Q. Sun, E.C. Chang, Detecting digital image forgeries by measuring inconsistencies of blocking artifact, in: Proceedings of the IEEE International Conference on Multimedia and Expo, IEEE, 2007, pp. 12–15.
[32] G.A. Triantafyllidis, D. Tzovaras, M.G. Strintzis, Detection of blocking artifacts of compressed still images, in: Proceedings of the 11th International Conference on Image Analysis and Processing, IEEE, 2001, pp. 607–611.
[33] A. Ignatov, K. Byeoung-su, R. Timofte, A. Pouget, Fast camera image denoising on mobile GPUS with deep learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 2515–2524.
[34] G. Algan, I. Ulusoy, Image classification with deep learning in the presence of noisy labels: a survey, Knowl. Based Syst. 215 (2021) 106771.
[35] S. J. Shin, S. C. You, H. Jeon, J. W. Jung, M. H. An, R. W. Park, J. Roh, Style transfer strategy for developing a generalizable deep learning application in digital pathology, Computer Methods and Programs in Biomedicine 198 (2021) 105815.
[36] X. Deng, Y. Zhang, M. Xu, S. Gu, Y. Duan, Deep coupled feedback network for joint exposure fusion and image super-resolution, IEEE Trans. Image Process. 30 (2021) 3098–3112.
[37] X. Zhou, B. Zhang, T. Zhang, P. Zhang, J. Bao, D. Chen, Z. Zhang, F. Wen, Cocosnet v2: Full-resolution correspondence learning for image translation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 11465–11475.
[38] L. Theis, W. Shi, A.C.F. Husz´ar, Lossy image compression with compressive autoencoders, Statistics 1050 (2017) 1.
[39] L. Zhou, C. Cai, Y. Gao, S. Su, J. Wu, Variational autoencoder for low bit-rate image compression, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018, pp. 2617–2620.
[40] Z. Cheng, H. Sun, M. Takeuchi, J. Katto, Deep convolutional autoencoderbased lossy image compression, in: Proceedings of the Picture Coding Symposium (PCS), IEEE, 2018, pp. 253–257.
[41] Y. Zhang, A better autoencoder for image: convolutional autoencoder„ in: Proceedings of the International Conference on Neural Information Processing ICONIP17-DCEC, 2018 Available online.
[42] Z. Cheng, H. Sun, M. Takeuchi, J. Katto, Performance comparison of convolutional autoencoders, generative adversarial networks and superresolution for image compression, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018, pp. 2613–2616.
[43] F. Xu, Z. Yan, G. Xiao, K. Zhang, W. Zuo, Jpeg image super-resolution via deep residual network, in: Proceedings of the International Conference on Intelligent Computing, Springer, 2018, pp. 472–483.
[44] T. Komatsu, Y. Ueda, T. Saito, Super-resolution decoding of jpegcompressed image data with the shrinkage in the redundant dct domain, in: Proceedings of the 28th Picture Coding Symposium, IEEE, 2010, pp. 114–117.
[45] M. Akbari, J. Liang, J. Han, Dsslic: deep semantic segmentation-based layered image compression, in: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2019, IEEE, 2019, pp. 2042–2046. ICASSP.
[46] S. Chen, S. Zhang, X. Zheng, X. Ruan, Layered adaptive compression design for efficient data collection in industrial wireless sensor networks, J. Netw. Comput. Appl. 129 (2019) 37–45 URL https://www.sciencedirect.com/science/article/pii/S1084804519300025 , doi:10.1016/j.jnca.2019.01.002.
[47] C. Jia, Z. Liu, Y. Wang, S. Ma, W. Gao, Layered image compression using scalable auto-encoder, in: Proceedings of the IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), 2019, pp. 431–436, doi:10.1109/MIPR.2019.00087.
[48] T.M. Hoang, J. Zhou, Y. Fan, Image compression with encoder-decoder matched semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 160–161.
[49] Y. Hu, S. Yang, W. Yang, L.Y. Duan, J. Liu, Towards coding for human and machine vision: a scalable image coding approach, in: Proceedings of the IEEE International Conference on Multimedia and Expo (ICME), IEEE, 2020, pp. 1–6.
[50] C. Wang, Y. Han, W. Wang, An end-to-end deep learning image compression framework based on semantic analysis, Appl. Sci. 9 (17) (2019) 3580.
[51] S. Luo, Y. Yang, Y. Yin, C. Shen, Y. Zhao, M. Song, Deepsic: deep semantic image compression, in: Proceedings of the International Conference on Neural Information Processing, Springer, 2018, pp. 96–106.
[52] L. Zhou, Z. Sun, X. Wu, J. Wu, End-to-end optimized image compression with attention mechanism, in: Proceedings of the Conference on Computer Vision and Pattern Recognition Workshops, 2019, p. 0.
[53] Z. Cheng, H. Sun, M. Takeuchi, J. Katto, Learned image compression with discretized gaussian mixture likelihoods and attention modules, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 7939–7948.
[54] N. Ouerhani, J. Bracamonte, H. Hugli, M. Ansorge, F. Pellandini, Adaptive color image compression based on visual attention, in: Proceedings of the 11th International Conference on Image Analysis and Processing, IEEE, 2001, pp. 416–421.
[55] H. Liu, T. Chen, Q. Shen, Z. Ma, Practical stacked non-local attention modules for image compression, in: Proceedings of the Conference on Computer Vision and Pattern Recognition Workshops, 2019, p. 0.
[56] G. Bjontegaard, Calculation of average psnr differences between rd-curves, VCEG-M33.
[57] Mean squared error, https://en.wikipedia.org/wiki/Mean_squared_error/, 2021, accessed 24 June 2021.
[58] J. Ball´e, V. Laparra, E. P. Simoncelli, End-to-end optimized image compression (2017). arXiv:1611.01704.
[59] G. Toderici, S. M. O'Malley, S. J. Hwang, D. Vincent, D. Minnen, S. Baluja, M. Covell, R. Sukthankar, Variable rate image compression with recurrent neural networks, arXiv preprint arXiv:1511.06085, 2015.
[60] G. Toderici, D. Vincent, N. Johnston, S.Jin Hwang, D. Minnen, J. Shor, M. Covell, Full resolution image compression with recurrent neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2017, pp. 5306–5314.
[61] F. Yang, L. Herranz, Y. Cheng, M.G. Mozerov, Slimmable compressive autoencoders for practical neural image compression, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 4998–5007.
[62] Z. Cheng, H. Sun, J. Katto, Low bitrate image compression with discretized gaussian mixture likelihoods, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, IEEE, 2020, pp. 126–127.
[63] H. Ma, D. Liu, N. Yan, H. Li, F. Wu, End-to-end optimized versatile image compression with wavelet-like transform, IEEE Transactions on Pattern Analysis and Machine Intelligence, IEEE, 2020.
[64] H. Ma, D. Liu, R. Xiong, F. Wu, Iwave: CNN-based wavelet-like transform for image compression, IEEE Transactions on Multimedia 22 (7) (2019) 1667–1679, doi:10.1109/TMM.2019.2957990.
[65] T. Chen, H. Liu, Z. Ma, Q. Shen, X. Cao, Y. Wang, End-to-end learnt image compression via non-local attention optimization and improved context modeling, IEEE Trans. Image Process. 30 (2021) 3179–3191.
[66] T. Dumas, A. Roumy, C. Guillemot, Autoencoder based image compression: can the learning be quantization independent? in: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2018, pp. 1188–1192. ICASSP.
[67] J. Ball´e, D. Minnen, S. Singh, S.J. Hwang, N. Johnston, Variational image compression with a scale hyperprior, in: Proceedings of the International Conference on Learning Representations, 2018.
[68] D. Minnen, G. Toderici, S. Singh, S.J. Hwang, M. Covell, Imagedependent local entropy models for learned image compression, in: Proceedings of the 25th IEEE International Conference on Image Processing, IEEE, 2018, pp. 430–434. (ICIP).
[69] D. Minnen, J. Ball´e, G. Toderici, Joint autoregressive and hierarchical priors for learned image compression, in: Proceedings of the 32nd International Conference on Neural Information Processing Systems, 2018, pp. 10794–10803.
[70] A.J. Hussain, D. Al-Jumeily, N. Radi, P. Lisboa, Hybrid neural network predictive-wavelet image compression system, Neurocomputing 151 (2015) 975–984.
[71] J. Klopp, Y.C.F. Wang, S.Y. Chien, L.G. Chen, Learning a code-space predictor by exploiting intra-image-dependencies, in: Proceedings of the British Machine Vision Conference, 2018, p. 124.

[72] J. Lee, S. Cho, S.K. Beack, Context-adaptive entropy model for end-toend optimized image compression, in: Proceedings of the International Conference on Learning Representations, 2018.
[73] F. Mentzer, E. Agustsson, M. Tschannen, R. Timofte, L. Van Gool, Conditional probability models for deep image compression, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4394–4402.
[74] N. Johnston, D. Vincent, D. Minnen, M. Covell, S. Singh, T. Chinen, S.J. Hwang, J. Shor, G. Toderici, Improved lossy image compression with priming and spatially adaptive bit rates for recurrent networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4385–4393.
[75] Y. Hu, W. Yang, J. Liu, Coarse-to-fine hyper-prior modeling for learned image compression, in: Proceedings of the AAAI Conference on Artificial Intelligence, 34, 2020, pp. 11013–11020. Vol..
[76] X. Deng, W. Yang, R. Yang, M. Xu, E. Liu, Q. Feng, R. Timofte, Deep homography for efficient stereo image compression, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 1492–1501.
[77] Y. Ge, J. Wang, Y. Shi, S. Gao, Hierarchical image compression framework, in: Neural Compression: From Information Theory to Applications– Workshop@ ICLR 2021, 2021.
[78] Y. Hu, W. Yang, Z. Ma, J. Liu, Learning end-to-end lossy image compression: a benchmark, IEEE Transactions on Pattern Analysis and Machine Intelligence (01) (2021) 1, doi:10.1109/TPAMI.2021.3065339.
[79] I. Schiopu, H. Huang, A. Munteanu, CNN-based intra-prediction for lossless hevc, IEEE Trans. Circuits Syst. Video Technol. 30 (7) (2019) 1816–1828.
[80] Z.T. Zhang, C.H. Yeh, L.W. Kang, M.H. Lin, Efficient ctu-based intra frame coding for hevc based on deep learning, in: Proceedings of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), IEEE, 2017, pp. 661–664.
[81] Y. Wang, X. Fan, S. Liu, D. Zhao, W. Gao, Multi-scale convolutional neural network-based intra prediction for video coding, IEEE Trans. Circuits Syst. Video Technol. 30 (7) (2019) 1803–1815.
[82] G. Zhong, J. Wang, J. Hu, F. Liang, A gan-based video intra coding, Electronics 10 (2) (2021) 132.
[83] L. Zhu, S. Kwong, Y. Zhang, S. Wang, X. Wang, Generative adversarial network-based intra prediction for video coding, IEEE Trans. Multimed. 22 (1) (2019) 45–58.
[84] Y. Hu, W. Yang, S. Xia, W.-H. Cheng, J. Liu, Enhanced intra prediction with recurrent neural network in video coding, in: inProceedings of the Data Compression Conference, IEEE, 2018, p. 413.
[85] Y. Hu, W. Yang, S. Xia, J. Liu, Optimized spatial recurrent network for intra prediction in video coding, in: Proceedings of the IEEE Visual Communications and Image Processing, IEEE, 2018, pp. 1–4. (VCIP).
[86] Y. Hu, W. Yang, M. Li, J. Liu, Progressive spatial recurrent neural network for intra prediction, IEEE Trans. Multimed. 21 (12) (2019) 3024–3037.
[87] W. Cui, T. Zhang, S. Zhang, F. Jiang, W. Zuo, Z. Wan, D. Zhao, Convolutional neural networks based intra prediction for hevc, in: Proceedings of the Data Compression Conference (DCC), IEEE Computer Society, 2017, p. 436.
[88] J. V. E. Team, Rec. itu-t h.265-iso/iec 23008-2 high efficiency video coding (hevc), https://vcgit.hhi.fraunhofer.de/jvet/HM, [Online; accessed 24-June- 2021].
[89] S. Ryu, J. Kang, Machine learning-based fast angular prediction mode decision technique in video coding, IEEE Transactions on Image Processing 27 (11) (2018) 5525–5538, doi:10.1109/TIP.2018.2857404.
[90] S. Kuanar, K. Rao, C. Conly, Fast mode decision in hevc intra prediction, using region wise cnn feature classification, in: Proceedings of the IEEE International Conference on Multimedia & Expo Workshops (ICMEW), IEEE, 2018, pp. 1–4.
[91] S. Kuanar, K. Rao, M. Bilas, J. Bredow, Adaptive cu mode selection in hevc intra prediction: a deep learning approach, Circuits Syst. Signal Process. 38 (11) (2019) 5081–5102.
[92] H. Lv, R. Wang, X. Xie, H. Jia, W. Gao, A comparison of fractional-pel interpolation filters in hevc and h. 264/avc, in: Proceedings of the Visual Communications and Image Processing, IEEE, 2012, pp. 1–6.
[93] A. Djelouah, J. Campos, S. Schaub-Meyer, C. Schroers, Neural interframe compression for video coding, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 6421–6429.
[94] J. Schneider, J. Sauer, M. Wien, Dictionary learning based high frequency inter-layer prediction for scalable hevc, in: Proceedings of the IEEE Visual Communications and Image Processing, IEEE, 2017, pp. 1–4. VCIP.
[95] J.-K. Lee, N. Kim, S. Cho, J.-W. Kang, Deep video prediction networkbased inter-frame coding in hevc, IEEE Access 8 (2020) 95906–95917.
[96] Y. Vatis, J. Ostermann, Adaptive interpolation filter for h. 264/avc, IEEE Trans. Circuits Syst. Video Technol. 19 (2) (2008) 179–192.
[97] H. Zhang, L. Song, Z. Luo, X. Yang, Learning a convolutional neural network for fractional interpolation in hevc inter coding, in: Proceedings of the IEEE Visual Communications and Image Processing, IEEE, 2017, pp. 1–4. VCIP.
[98] N. Yan, D. Liu, H. Li, T. Xu, F. Wu, B. Li, Convolutional neural networkbased invertible half-pixel interpolation filter for video coding, in: Proceedings of the 25th IEEE International Conference on Image Processing (ICIP), IEEE, 2018, pp. 201–205.
[99] C.D.K. Pham, J. Zhou, Deep learning-based luma and chroma fractional interpolation in video coding, IEEE Access 7 (2019) 112535–112543.
[100] J.K. Lee, N. Kim, S. Cho, J.-W. Kang, Convolution neural network based video coding technique using reference video synthesis, in: Proceedings of the Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), IEEE, 2018, pp. 505–508.
[101] L. Zhao, S. Wang, X. Zhang, S. Wang, S. Ma, W. Gao, Enhanced ctulevel inter prediction with deep frame rate up-conversion for high efficiency video coding, in: Proceedings of the 25th IEEE International Conference on Image Processing, IEEE, 2018, pp. 206–210. (ICIP).
[102] L. Zhao, S. Wang, X. Zhang, S. Wang, S. Ma, W. Gao, Enhanced motioncompensated video coding with deep virtual reference frame generation, IEEE Trans. Image Process. 28 (10) (2019) 4832–4844.
[103] M. Benjak, H. Meuel, T. Laude, J. Ostermann, Enhanced machine learning-based inter coding for vvc, in: Proceedings of the International Conference on Artificial Intelligence in Information and Communication (ICAIIC), IEEE, 2021, pp. 021–025.
[104] Z. Zhao, S. Wang, S. Wang, X. Zhang, S. Ma, J. Yang, Cnn-based bidirectional motion compensation for high efficiency video coding, in: Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS), IEEE, 2018, pp. 1–4.
[105] Z. Zhao, S. Wang, S. Wang, X. Zhang, S. Ma, J. Yang, Enhanced biprediction with convolutional neural network for high-efficiency video coding, IEEE Trans. Circuits Syst. Video Technol. 29 (11) (2018) 3291–3301.
[106] G. Cote, B. Erol, M. Gallant, F. Kossentini, H. 263+: Video coding at low bit rates, IEEE Trans. Circuits Syst. Video Technol. 8 (7) (1998) 849–866.
[107] A. Norkin, G. Bjontegaard, A. Fuldseth, M. Narroschke, M. Ikeda, K. Andersson, M. Zhou, G. Van der Auwera, HEVC deblocking filter, IEEE Trans. Circuits Syst. Video Technol. 22 (12) (2012) 1746–1754.
[108] C.M. Fu, E. Alshina, A. Alshin, Y.W. Huang, C.Y. Chen, C.Y. Tsai, C.W. Hsu, S.M. Lei, J.H. Park, W.J. Han, Sample adaptive offset in the hevc standard, IEEE Trans. Circuits Syst. Video Technol. 22 (12) (2012) 1755–1764.
[109] C.Y. Tsai, C.Y. Chen, T. Yamakage, I.S. Chong, Y.W. Huang, C.M. Fu, T. Itoh, T. Watanabe, T. Chujoh, M. Karczewicz, et al., Adaptive loop filtering for video coding, IEEE J. Sel. Top. Signal Process. 7 (6) (2013) 934–945.
[110] X. Zhang, R. Xiong, W. Lin, J. Zhang, S. Wang, S. Ma, W. Gao, Lowrank-based nonlocal adaptive loop filter for high-efficiency video compression, IEEE Trans. Circuits Syst. Video Technol. 27 (10) (2016) 2177–2188.
[111] S. Ma, X. Zhang, J. Zhang, C. Jia, S. Wang, W. Gao, Nonlocal in-loop filter: the way toward next-generation video coding? IEEE Multimed. 23 (2) (2016) 16–26.
[112] W.-S. Park, M. Kim, CNN-based in-loop filtering for coding efficiency improvement, in: Proceedings of the IEEE 12th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP), 2016, pp. 1–5, doi:10.1109/IVMSPW. 2016.7528223.
[113] Y. Wang, H. Zhu, Y. Li, Z. Chen, S. Liu, Dense residual convolutional neural network based in-loop filter for hevc, in: Proceedings of the IEEE Visual Communications and Image Processing, 2018, pp. 1–4, doi:10.1109/VCIP.2018.8698740. VCIP.
[114] S. Kuanar, C. Conly, K. Rao, Deep learning based hevc in-loop filtering for decoder quality enhancement, in: Proceedings of the Picture Coding Symposium (PCS), IEEE, 2018, pp. 164–168.

[115] Y. Zhang, T. Shen, X. Ji, Y. Zhang, R. Xiong, Q. Dai, Residual highway convolutional neural networks for in-loop filtering in hevc, IEEE Trans. Image Process. 27 (8) (2018) 3827–3841.
[116] H. Huang, I. Schiopu, A. Munteanu, Frame-wise cnn-based filtering for intra-frame quality enhancement of hevc videos, IEEE Trans. Circuits Syst. Video Technol. 2020 (2020) TCSVT–2020.
[117] C. Sun, X. Fan, D. Zhao, Deep learning based intra prediction filter in avs3, in: Proceedings of the IEEE International Conference on Multimedia & Expo Workshops (ICMEW), IEEE, 2020, pp. 1–6.
[118] C. Jia, S. Wang, X. Zhang, S. Wang, S. Ma, Spatial-temporal residue network based in-loop filter for video coding, in: Proceedings of the IEEE Visual Communications and Image Processing, IEEE, 2017, pp. 1–4. VCIP.
[119] X. Meng, C. Chen, S. Zhu, B. Zeng, A new hevc in-loop filter based on multi-channel long-short-term dependency residual networks, in: Proceedings of the Data Compression Conference, IEEE, 2018, pp. 187–196.
[120] T. Li, M. Xu, C. Zhu, R. Yang, Z. Wang, Z. Guan, A deep learning approach for multi-frame in-loop filter of hevc, IEEE Trans. Image Process. 28 (11) (2019) 5663–5678.
[121] T. Li, M. Xu, R. Yang, X. Tao, A densenet based approach for multiframe in-loop filter in hevc, in: Proceedings of the Data Compression Conference (DCC), IEEE, 2019, pp. 270–279.
[122] Z. Wang, C. Ma, R.-L. Liao, Y. Ye, Multi-density convolutional neural network for in-loop filter in video coding, in: Proceedings of the Data Compression Conference (DCC), IEEE, 2021, pp. 23–32.
[123] J. V. E. Team, Rec. itu-t h.266 — iso/iec 23090-3 versatile video coding (vvc), https://vcgit.hhi.fraunhofer.de/jvet/VVCSoftware_VTM, 2021, accessed 24 June 2021.
[124] X. Song, J. Yao, L. Zhou, L. Wang, X. Wu, D. Xie, S. Pu, A practical convolutional neural network as loop filter for intra frame, in: Proceedings of the 25th IEEE International Conference on Image Processing (ICIP), IEEE, 2018, pp. 1133–1137.
[125] Z. Huang, X. Guo, M. Shang, J. Gao, J. Sun, An efficient qp variable convolutional neural network based in-loop filter for intra coding, in: Proceedings of the Data Compression Conference (DCC), IEEE, 2021, pp. 33–42.
[126] C. Lan, J. Xu, F. Wu, Improving depth compression in hevc by pre/post processing, in: Proceedings of the IEEE International Conference on Multimedia and Expo Workshops, IEEE, 2012, pp. 611–616.
[127] Y. Dai, D. Liu, F. Wu, A convolutional neural network approach for postprocessing in hevc intra coding, in: Proceedings of the International Conference on Multimedia Modeling, Springer, 2017, pp. 28–39.
[128] C. Li, L. Song, R. Xie, W. Zhang, CNN based post-processing to improve hevc, in: Proceedings of the IEEE International Conference on Image Processing (ICIP), IEEE, 2017, pp. 4577–4580.
[129] L. Ma, Y. Tian, T. Huang, Residual-based video restoration for hevc intra coding, in: Proceedings of the IEEE Fourth International Conference on Multimedia Big Data (BigMM), IEEE, 2018, pp. 1–7.
[130] F. Li, W. Tan, B. Yan, Deep residual network for enhancing quality of the decoded intra frames of hevc, in: Proceedings of the 25th IEEE International Conference on Image Processing (ICIP), IEEE, 2018, pp. 3918–3922.
[131] D. Ma, F. Zhang, D.R. Bull, Mfrnet: a new cnn architecture for postprocessing and in-loop filtering, IEEE J. Sel. Top. Signal Process. 15 (2) (2020) 378–387.
[132] F. Zhang, C. Feng, D.R. Bull, Enhancing VVC through cnn-based postprocessing, in: Proceedings of the IEEE International Conference on Multimedia and Expo (ICME), IEEE, 2020, pp. 1–6.
[133] P.R. Lai, J.S. Wang, Multi-stage attention convolutional neural networks for hevc in-loop filtering, in: Proceedings of the 2nd IEEE International Conference on Artificial Intelligence Circuits and Systems (AICAS), IEEE, 2020, pp. 173–177.
[134] W. Lin, X. He, X. Han, D. Liu, J. See, J. Zou, H. Xiong, F. Wu, Partitionaware adaptive switching neural networks for post-processing in hevc, IEEE Trans. Multimed. 22 (11) (2019) 2749–2763.
[135] X. He, Q. Hu, X. Zhang, C. Zhang, W. Lin, X. Han, Enhancing hevc compressed videos with a partition-masked convolutional neural network, in: Proceedings of the 25th IEEE International Conference on Image Processing (ICIP), IEEE, 2018, pp. 216–220.
[136] Z. Guan, Q. Xing, M. Xu, R. Yang, T. Liu, Z. Wang, Mfqe 2.0: a new approach for multi-frame quality enhancement on compressed video, IEEE Trans. Pattern Anal. Mach. Intell. 43 (03) (2021) 949–963.
[137] J. Tong, X. Wu, D. Ding, Z. Zhu, Z. Liu, Learning-based multi-frame video quality enhancement, in: Proceedings of the IEEE International Conference on Image Processing (ICIP), IEEE, 2019, pp. 929–933.
[138] M. Lu, M. Cheng, Y. Xu, S. Pu, Q. Shen, Z. Ma, Learned quality enhancement via multi-frame priors for hevc compliant low-delay applications, in: Proceedings of the IEEE International Conference on Image Processing, IEEE, 2019, pp. 934–938. ICIP.
[139] J. Deng, L. Wang, S. Pu, C. Zhuo, Spatio-temporal deformable convolution for compressed video quality enhancement, in: Proceedings of the AAAI Conference on Artificial Intelligence, 34, 2020, pp. 10696–10703. Vol..
[140] Y. Xu, M. Zhao, J. Liu, X. Zhang, L. Gao, S. Zhou, H. Sun, Boosting the performance of video compression artifact reduction with reference frame proposals and frequency domain information, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 213–222.
[141] T.M. Hoang, J. Zhou, B-drrn: A block information constrained deep recursive residual network for video compression artifacts reduction, in: Proceedings of the Picture Coding Symposium (PCS), IEEE, 2019, pp. 1–5.
[142] O. Rippel, S. Nair, C. Lew, S. Branson, A.G. Anderson, L. Bourdev, Learned video compression, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 3454–3463.
[143] M. Shen, P. Xue, C. Wang, Down-sampling based video coding using super-resolution technique, IEEE Trans. Circuits Syst. Video Technol. 21 (6) (2011) 755–765.
[144] M. Afonso, F. Zhang, A. Katsenou, D. Agrafiotis, D. Bull, Low complexity video coding based on spatial resolution adaptation, in: Proceedings of the IEEE International Conference on Image Processing (ICIP), IEEE, 2017, pp. 3011–3015.
[145] H. Lin, X. He, L. Qing, Q. Teng, S. Yang, Improved low-bitrate hevc video coding using deep learning based super-resolution and adaptive block patching, IEEE Trans. Multimed. 21 (12) (2019) 3010–3023.
[146] F. Nasiri, W. Hamidouche, L. Morin, G. Cocherel, N. Dhollande, A study on the impact of training data in cnn-based super-resolution for low bitrate end-to-end video coding, in: Proceedings of the Tenth International Conference on Image Processing Theory, Tools and Applications (IPTA), IEEE, 2020, pp. 1–5.
[147] D. Liu, Z. Chen, S. Liu, F. Wu, Deep learning-based technology in responses to the joint call for proposals on video compression with capability beyond hevc, IEEE Trans. Circuits Syst. Video Technol. 30 (5) (2019) 1267–1280.
[148] M.M. Ho, J. Zhou, G. He, M. Li, L. Li, SR-CL-DMC: P-frame coding with super-resolution, color learning, and deep motion compensation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 124–125.
[149] F. Li, H. Bai, Y. Zhao, Learning a deep dual attention network for video super-resolution, IEEE Trans. Image Process. 29 (2020) 4474–4488.
[150] F. Zhang, M. Afonso, D.R. Bull, Vistra2: video coding using spatial resolution and effective bit depth adaptation, Signal Processing: Image Communication, 97 (2021) 116355 116355, doi:10.1016/j.image.2021.116355.
[151] G. He, C. Wu, L. Li, J. Zhou, X. Wang, Y. Zheng, B. Yu, W. Xie, A video compression framework using an overfitted restoration neural network, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 148–149.
[152] Y. Li, D. Liu, H. Li, L. Li, F. Wu, H. Zhang, H. Yang, Convolutional neural network-based block up-sampling for intra frame coding, IEEE Trans. Circuits Syst. Video Technol. 28 (9) (2017) 2316–2330.
[153] J. Lin, D. Liu, H. Yang, H. Li, F. Wu, Convolutional neural network-based block up-sampling for hevc, IEEE Trans. Circuits Syst. Video Technol. 29 (12) (2018) 3701–3715.
[154] D. Ma, F. Zhang, D.R. Bull, Gan-based effective bit depth adaptation for perceptual video compression, in: Proceedings of the IEEE International Conference on Multimedia and Expo (ICME), IEEE, 2020, pp. 1–6.
[155] Z. Pan, F. Yuan, J. Lei, S. Kwong, Video compression coding via colorization: a generative adversarial network (gan)-based approach, arXiv preprint arXiv:1912.10653, 2019.

[156] S. Kim, J.S. Park, C.G. Bampis, J. Lee, M.K. Markey, A.G. Dimakis, A.C. Bovik, Adversarial video compression guided by soft edge detection, in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2020, IEEE, 2020, pp. 2193–2197. ICASSP.
[157] X. Sun, X. Yang, S. Wang, M. Liu, Content-aware rate control scheme for hevc based on static and dynamic saliency detection, Neurocomputing 411 (2020) 393–405.
[158] H. Ko, H.Y. Kim, Deep learning-based compression for phase-only hologram, IEEE Access 9 (2021) 79735–79751 3084800., doi:10.1109/ACCESS.2021.
[159] Z. Chen, T. He, Learning based facial image compression with semantic fidelity metric, Neurocomputing 338 (2019) 16–25.
[160] L. Galteri, M. Bertini, L. Seidenari, T. Uricchio, A. Del Bimbo, Increasing video perceptual quality with gans and semantic coding, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 862–870.
[161] X. Li, J. Shi, Z. Chen, Task-driven semantic coding via reinforcement learning, arXiv preprint arXiv:2106.03511, 2021.
[162] A. Hassan, M. Ghafoor, S.A. Tariq, T. Zia, W. Ahmad, High efficiency video coding (HEVC)-based surgical telementoring system using shallow convolutional neural network, J. Digit. Imaging 32 (6) (2019) 1027–1043.
[163] D. Ding, J. Tong, L. Kong, A deep learning approach for quality enhancement of surveillance video, J. Intell. Transp. Syst. 24 (3) (2020) 304–314.
[164] J. Zhang, C. Jia, M. Lei, S. Wang, S. Ma, W. Gao, Recent development of avs video coding standard: Avs3, in: Proceedings of the Picture Coding Symposium (PCS), IEEE, 2019, pp. 1–5.
[165] C. Jia, S. Wang, X. Zhang, S. Wang, J. Liu, S. Pu, S. Ma, Content-aware convolutional neural network for in-loop filtering in high efficiency video coding, IEEE Trans. Image Process. 28 (7) (2019) 3343–3356.
[166] H. Huang, I. Schiopu, A. Munteanu, Low-complexity angular intraprediction convolutional neural network for lossless hevc, in: Proceedings of the IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP), IEEE, 2020, pp. 1–6.
[167] W. Kuang, Y.L. Chan, S.H. Tsang, Low-complexity intra prediction for screen content coding by convolutional neural network, in: Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS), IEEE, 2020, pp. 1–5.
[168] C. Liu, H. Sun, J. Katto, X. Zeng, Y. Fan, A learning-based low complexity in-loop filter for video coding, in: Proceedings of the IEEE International Conference on Multimedia & Expo Workshops (ICMEW), IEEE, 2020, pp. 1–6.
[169] T. Li, M. Xu, X. Deng, A deep convolutional neural network approach for complexity reduction on intra-mode HEVC, in: Proceedings of the IEEE International Conference on Multimedia and Expo (ICME), IEEE, 2017, pp. 1255–1260.
[170] S. Bouaafia, R. Khemiri, F.E. Sayadi, M. Atri, Fast cu partition-based machine learning approach for reducing hevc complexity, J. Real Time Image Process. 17 (1) (2020) 185–196.
[171] C. Huang, Z. Peng, Y. Xu, F. Chen, Q. Jiang, Y. Zhang, G. Jiang, Y.-S. Ho, Online learning-based multi-stage complexity control for live video coding, IEEE Trans. Image Process. 30 (2020) 641–656.
[172] L. Yu, L. Shen, H. Yang, X. Jiang, B. Yan, A distortion-aware multitask learning framework for fractional interpolation in video coding, IEEE Trans. Circuits Syst. Video Technol. 31 (7) (2021) 2824–2836, doi:10.1109/TCSVT.2020.3028330.
[173] C. Liu, H. Sun, J. Katto, X. Zeng, Y. Fan, A qp-adaptive mechanism for cnn-based filter in video coding, arXiv preprint arXiv:2010.13059, 2020.
[174] G. Lu, W. Ouyang, D. Xu, X. Zhang, C. Cai, Z. Gao, Dvc: An end-to-end deep video compression framework, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 11006–11015.
[175] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Van Der, D. Smagt, T. Cremers, Brox, Flownet: learning optical flow with convolutional networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 2758–2766.
[176] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, T. Brox, Flownet 2.0: evolution of optical flow estimation with deep networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2462–2470.
[177] E. Agustsson, D. Minnen, N. Johnston, J. Balle, S.J. Hwang, G. Toderici, Scale-space flow for end-to-end optimized video compression, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 8503–8512.
[178] L. Wang, Y. Guo, L. Liu, Z. Lin, X. Deng, W. An, Deep video superresolution using hr optical flow estimation, IEEE Trans. Image Process. 29 (2020) 4323–4336.
[179] A. Ranjan, M.J. Black, Optical flow estimation using a spatial pyramid network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4161–4170.
[180] A. Nosratinia, M.T. Orchard, Optimal warping prediction for video coding, in: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings, Vol. 4, IEEE, 1996, pp. 1986–1989.
[181] M. Wang, G.Y. Yang, J.K. Lin, S.H. Zhang, A. Shamir, S.P. Lu, S.M. Hu, Deep online video stabilization with multi-grid warping transformation learning, IEEE Trans. Image Process. 28 (5) (2018) 2283–2292.
[182] M. Zhao, Q. Ling, Pwstablenet: learning pixel-wise warping maps for video stabilization, IEEE Trans. Image Process. 29 (2020) 3582–3595.
[183] T.C. Wang, M.Y. Liu, J.Y. Zhu, A. Tao, J. Kautz, B. Catanzaro, Highresolution image synthesis and semantic manipulation with conditional gans, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 8798–8807.
[184] Q. Chen, V. Koltun, Photographic image synthesis with cascaded refinement networks, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 1511–1520.
[185] C. Jia, X. Zhang, S. Wang, S. Wang, S. Ma, Light field image compression using generative adversarial network-based view synthesis, IEEE J. Emerg. Sel. Top. Circuits Syst. 9 (1) (2018) 177–189.
[186] J. Lin, D. Liu, H. Li, F. Wu, M-lvc: multiple frames prediction for learned video compression, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 3546–3554.
[187] X. Meng, X. Deng, S. Zhu, X. Zhang, B. Zeng, A robust quality enhancement method based on joint spatial-temporal priors for video coding, IEEE Trans. Circuits Syst. Video Technol. 31 (6) (2021) 2401–2414, doi:10.1109/TCSVT.2020.3019919.
[188] A. Golinski, R. Pourreza, Y. Yang, G. Sautiere, T.S. Cohen, Feedback recurrent autoencoder for video compression, in: Proceedings of the Asian Conference on Computer Vision, 2020.
[189] R. Feng, Y. Wu, Z. Guo, Z. Zhang, Z. Chen, Learned video compression with feature-level residuals, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 120–121.
[190] Z. Hu, G. Lu, D. Xu, Fvc: A new framework towards deep video compression in feature space, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 1502–1511.
[191] R. Yang, F. Mentzer, L.V. Gool, R. Timofte, Learning for video compression with hierarchical quality and recurrent enhancement, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 6628–6637.
[192] Z. Hu, Z. Chen, D. Xu, G. Lu, W. Ouyang, S. Gu, Improving deep video compression by resolution-adaptive flow coding, in: Proceedings of the European Conference on Computer Vision, Springer, 2020, pp. 193–209.
[193] Z. Cheng, H. Sun, M. Takeuchi, J. Katto, Learning image and video compression through spatial-temporal energy compaction, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 10071–10080.
[194] O. Rippel, A. G. Anderson, K. Tatwawadi, S. Nair, C. Lytle, L. Bourdev, Elf-vc: Efficient learned flexible-rate video coding, arXiv preprint, arXiv:2104.14335, 2021.
[195] Z. Guo, Z. Zhang, R. Feng, Z. Chen, Soft then hard: rethinking the quantization in neural image compression, arXiv preprint arXiv:2104.05168, 2021.
[196] R. Yang, Y. Yang, J. Marino, S. Mandt, Hierarchical autoregressive modeling for neural video compression, in: Proceedings of the International Conference on Learning Representations, 2020.
[197] A. Habibian, T.v. Rozendaal, J.M. Tomczak, T.S. Cohen, Video compression with rate-distortion autoencoders, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 7033–7042.
[198] J. Han, S. Lombardo, C. Schroers, S. Mandt, Deep generative (2018).
[199] J. Luo, S. Li, W. Dai, Y. Xu, D. Cheng, G. Li, H. Xiong, Noise-tocompression variational autoencoder for efficient end-to-end optimized image coding, in: Proceedings of the Data Compression Conference (DCC), IEEE, 2020, pp. 33–42.

[200] N.R. Goodman, Statistical analysis based on a certain multivariate complex gaussian distribution (an introduction), Ann. Math. Stat. 34 (1) (1963) 152–177.

[201] N. Zou, H. Zhang, F. Cricri, H.R. Tavakoli, J. Lainema, E. Aksu, M. Hannuksela, E. Rahtu, End-to-end learning for video frame compression with self-attention, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 142–143.

[202] R. Yang, F. Mentzer, L. Van Gool, R. Timofte, Learning for video compression with recurrent auto-encoder and recurrent probability model, IEEE J. Sel. Top. Signal Process. 15 (2) (2021) 388–401, doi:10.1109/JSTSP.2020.3043590.

[203] C. Lin, J. Yao, F. Chen, L. Wang, A spatial rnn codec for end-to-end image compression, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 13269–13277.

[204] B. Liu, Y. Chen, S. Liu, H.-S. Kim, Deep learning in latent space for video prediction and compression, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 701–710.

[205] V. Veerabadran, R. Pourreza, A. Habibian, T.S. Cohen, Adversarial distortion for learned video compression, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 168–169.

[206] Z. Wang, E.P. Simoncelli, A.C. Bovik, Multiscale structural similarity for image quality assessment, in: Proceedings of the Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2, Ieee, 2003, pp. 1398–1402.

[207] A. Mercat, M. Viitanen, J. Vanne, Uvg dataset: 50/120fps 4k sequences for video codec analysis and development, in: Proceedings of the 11th ACM Multimedia Systems Conference, 2020, pp. 297–302.

[208] H. Wang, W. Gan, S. Hu, J.Y. Lin, L. Jin, L. Song, P. Wang, I. Katsavounidis, A. Aaron, C.-C.J. Kuo, Mcl-jcv: a jnd-based h. 264/avc video quality assessment dataset, in: Proceedings of the IEEE International Conference on Image Processing, IEEE, 2016, pp. 1509–1513. ICIP.

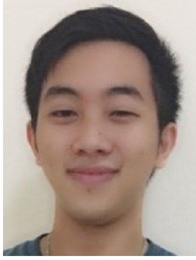
[209] Arizona State University, Video trace library, http://trace.eas.asu.edu/, 2021, accessed 30 June 2021.

[210] G. Sullivan, Common. http://phenix.it-sudparis.eu/jct/doc_end_user/current_document.php?id=7281/.

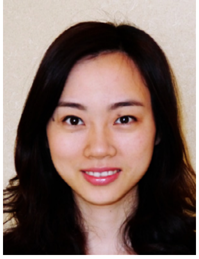
[211] R. Zhang, P. Isola, A.A. Efros, E. Shechtman, O. Wang, The unreasonable effectiveness of deep features as a perceptual metric, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 586–595.

[212] M Heusel, H Ramsauer, T Unterthiner, B Nessler, S Hochreiter, GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium, Advances in Neural Information Processing Systems 30 (2017).

[213] M. Bin´kowski, D.J. Sutherland, M. Arbel, A. Gretton, Demystifying mmd gans, in: Proceedings of the International Conference on Learning Representations, 2018.



Trinh Man Hoang: received his B.E. degree (Honors) in Computer Science from the University of Information Technology, Vietnam National University in 2018, M.E. degree in Apllied Informatics from Hosei University, Japan in 2020. Currently, he is a PhD student and a member of the Intelligent Multimedia Processing Lab (iMedia Lab) at Hosei University, Tokyo, Japan. His current research interests include Video Coding, Deep Learning, and Multimedia Processing.



Jinjia Zhou: (S'12-M'13) received B.E. degree from Shanghai Jiao Tong University, China, in 2007. She received M.E. and Ph.D. degrees from Waseda University, Japan, in 2010 and 2013, respectively. From 2013 to 2016, she was a junior researcher with Waseda University, Fukuoka, Japan. Currently, she is an Associate Professor and a co-director of the English based graduate program at Hosei University.She is also a senior visiting scholar in State Key Laboratory of ASIC & System, Fudan University, China. From 2020, She is also a specially appointed Assoc. Prof. with Osaka University. Her interests are in algorithms and VLSI architectures for multimedia signal processing. Dr. Zhou was selected as JST PRESTO researcher during 2017-2021. She received the research fellowship of the Japan Society for the Promotion of Science during 2010-2013.She received the Hibikino Best Thesis Award in 2011. She was a co-recipient of the Best Paper Runner-up Award of IEEE MMM 2020, IEEE ISSCC 2016 Takuo Sugano Award for Outstanding Far-East Paper, the best student paper award of VLSI Circuits Symposium 2010 and the design contest award of ACM ISLPED 2010. She participated in the design of the world first 8K UHDTV video decoder chip, which was granted the 2012 Semiconductor of the Year Award of Japan.