Survey on Visual Signal Coding and Processing With Generative Models: Technologies, Standards, and Optimization

Zhibo Chen, *Senior Member, IEEE*, Heming Sun, *Member, IEEE*, Li Zhang, *Senior Member, IEEE*, and Fan Zhang, *Member, IEEE*

***Abstract*— This paper provides a survey of the latest developments in visual signal coding and processing with generative models. Specifically, our focus is on presenting the advancement of generative models and their influence on research in the domain of visual signal coding and processing. This survey study begins with a brief introduction of well-established generative models, including the Variational Autoencoder (VAE) models, Generative Adversarial Network (GAN) models, Autoregressive (AR) models, Normalizing Flows and Diffusion models. The subsequent section of the paper explores the advancements in visual signal coding based on generative models, as well as the ongoing international standardization activities. In the realm of visual signal processing, our focus lies on the application and development of various generative models in the research of visual signal restoration. We also present the latest developments in generative visual signal synthesis and editing, along with visual signal quality assessment using generative models and quality assessment for generative models. The practical implementation of these studies is closely linked to the investigation of fast optimization. This paper additionally presents the latest advancements in fast optimization on visual signal coding and processing with generative models. We hope to advance this field by providing researchers and practitioners a comprehensive literature review on the topic of visual signal coding and processing with generative models.**

***Index Terms*— Generative models, visual signal coding, visual signal processing, optimization.**

I. Introduction

IN RECENT years, generative models have emerged as one of the most significant and rapidly developing areas of research in artificial intelligence. Generative models have demonstrated remarkable success in synthesizing high-quality data (text, image, video, 3D content, etc.) and hold promise for utilizing unlabeled data, transfer learning, data augmentation, drug/protein generation, and other applications. Different generative models have been developed to capture complex data distributions and generate new examples. For example, the Autoregressive (AR) models, like GPT-4, sequentially predict and generate data. The Variational Autoencoders (VAE) learn the parameters of a probability distribution representing the input data. The Generative Adversarial Networks (GAN) train two competing neural networks in an adversarial manner to generate realistic synthetic data. The flow models learn invertible mappings between data and latent space. The Diffusion models iteratively add noise to data and then train a neural network to reverse the process.

Simultaneously, generative models have been demonstrated to be a crucial tool for learning-based visual signal coding and processing. For example, the VAE model has been widely employed as a foundational framework in end-to-end learning-based image coding schemes. The AR model has been extensively studied to improve entropy coding performance, and the GAN model and Diffusion model have been utilized frequently to enhance the subjective quality of coding schemes. Additionally, generative models have also been explored in various visual signal processing tasks, including restoration, enhancement, editing, quality assessment, and interpolation.

In light of the rapid growth of visual signal coding and processing with generative models, its contributions to international standards and practical application optimization are increasingly valued. The Joint Video Experts Team (JVET) of International Telecommunication Union - Telecommunication Sector Video Coding Experts Group (ITU-T VCEG) and International Organization for Standardization/International Electrotechnical Commission Moving Picture Experts Group (ISO/IEC MPEG) has started working together on an exploration study to evaluate potential neural network-based video coding (NNVC) technology beyond the capabilities of the conventional hybrid video coding framework as early as 2018. In addition, MPEG also launched many standardization projects, which have started adopting artificial intelligence (AI)-based technologies, such as AI-based 3D graphics coding, AI model compression, and video coding for machines (VCM). Meanwhile, the Joint Photographic Experts Group

Manuscript received 14 February 2024; revised 17 April 2024 and 16 May 2024; accepted 16 May 2024. Date of publication 21 May 2024; date of current version 27 June 2024. This work was supported in part by NSFC under Grant 62371434 and Grant 62021001, in part by JSPS KAKENHI under Grant JP23K16861, and in part by the U.K. Research and Innovation (UKRI) MyWorld Strength in Places Program under Grant SIPF00006/1. This article was recommended by Guest Editor W.-H. Peng. *(Corresponding author: Zhibo Chen.)*

Zhibo Chen is with the Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei 230026, China (e-mail: chenzhibo@ustc.edu.cn).

Heming Sun is with the Faculty of Engineering, Yokohama National University, Kanagawa 240-0067, Japan (e-mail: sun-heming-vg@ynu.ac.jp).

Li Zhang is with Bytedance Inc., San Diego, CA 92122 USA (e-mail: lizhang.idm@bytedance.com).

Fan Zhang is with the School of Computer Science, University of Bristol, BS8 1UB Bristol, U.K. (e-mail: fan.zhang@bristol.ac.uk).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/JETCAS.2024.3403524.

Digital Object Identifier 10.1109/JETCAS.2024.3403524

(JPEG) AI, the creation of a learning-based image coding standard and a joint standardization project between ISO/IEC JPEG and ITU-T, is being developed. It is observed that these standards built on AI-based methods are able to serve various purposes, including significant compression efficiency improvement, efficient multimedia representation, compression and deployment of neural network models, metadata extraction for AI-based processes, and multimedia content processing. Although learning-based visual signal coding has obtained a remarkable coding gain, the complexity is still too high to achieve real-time processing. Regarding the algorithm, the AR model is usually used to fully utilize the correlation of neighboring samples. In addition, the huge scale of the neural network increases the burden of computational and memory cost. Regarding hardware and implementation, most frameworks are accelerated by general-purpose accelerators such as Nvidia GPU and AMD/Xilinx DPU. However, the hardware utilization of GPU/DPU is relatively low since the computation-to-communication (CTC) ratio of neural networks may not fit the hardware resources such as the number of PE cores and bandwidth. Recently, transformer-like architectures have gradually become popular due to their superior performance. To accommodate the rapid development of neural network architectures, developing a generic and efficient hardware accelerator is also a challenge.

Therefore, we seek to further advance this important research area by providing researchers and practitioners with a broad reference to the literature on visual signal coding and processing with generative models. The remainder of this paper provides a brief overview of the visual signal coding and processing with generative models from the perspectives of the background of generative models, the development of visual signal coding with generative models and related standardization activities, and the recent progress of visual signal processing with generative models. We start with an overview of generative models in Section II and then introduce visual signal coding with generative models in Section III, with an introduction of standardization activities on visual signal coding with generative models in Section IV. In Section V, we review some ongoing research activities related to visual signal processing with generative models, including restoration, enhancement, editing, and interpolation. The quality assessment methods based on generative models and those for generative models are further discussed in Section VI. Finally, a review of fast implementation and optimization of visual signal coding and processing with generative models is presented in Section VII, followed by a conclusion in Section VIII.

II. Generative Models

This section provides a brief introduction to generative models, including GANs, VAEs, Autoregressive Models, Normalizing Flows, and Diffusion models.

A. Generative Adversarial Networks (GAN)

GANs [1] are important components of deep generative models. They are developed to generate data through an adversarial training strategy involving both generators and discriminators. In this context, the objective of the generator is to generate data samples that are as realistic as possible to deceive the discriminator. The discriminator is tasked with differentiating these generated data samples from the actual ones in the training data, and both networks are updated iteratively. The objective is to find an equilibrium point where the discriminator cannot reliably discern fake data samples from real data samples. Utilizing the discriminator's ability to understand perception, GANs achieve outstanding results in qualitative generation and are widely used in a wide range of tasks, including conditional generation [2], representation learning [3], image-to-image translation [4], image super-resolution [5], image enhancement [6], style transfer [7] and semantic editing [8]. GANs are also capable of generalizing to generate data in modalities other than images, such as video [9], text [10], audio [11] and 3D data [12]. However, adversarial training can be difficult due to issues such as instability and mode collapse [13], which lead to low-quality generation outputs with limited variability. Some research [14] has attempted to mitigate this issue by employing more rational loss functions.

B. Variational Autoencoders (VAE)

VAEs [15] are a type of generative model that employ Bayesian inference to approximate the distribution of data. VAEs comprise two components: an encoder that transforms input data into a distribution of latent variables, and a decoder that reconstructs input data from the latent variables. The training objective of VAEs is to optimize a lower bound on the data log-likelihood, which is composed of two terms: a Kullback–Leibler (KL) divergence term that quantifies the dissimilarity between the distribution of latent variables and the prior distribution, and a reconstruction term that measures the fidelity of the generated data. VAEs have been improved by addressing several challenges, such as increasing the expressiveness of the latent variable distribution [16], reducing the gap between the lower bound and the true log-likelihood, avoiding posterior collapse [17], and scaling up to high-resolution data [18]. Some of the notable works that have contributed to these improvements are IAF-VAE [19], NVAE [18], and VDVAE [20], which have enhanced the performance and quality of VAEs, making them more powerful for generative modeling. Meanwhile, VAEs are increasingly used as part of other generative models, such as normalizing flows and diffusion models. These combinations have the added benefit of enhancing the performance of the generated data samples.

C. Autoregressive Models

Autoregressive models [21] view generation as a sequential process, predicting future outcomes based on previously observed data. They excel in precision, optimizing the likelihood of the estimated data by learning dependencies within the sequence. This is typically achieved through a masking strategy, such as PixelRNN [22] and Gated PixelCNN [23], where certain known values are used to predict unknown neighboring values. Autoregressive models exhibit exceptional

performance in modeling distribution density and capturing intricate patterns in data. However, their sequential nature results in a slow data generation process [24]. Furthermore, by relying on historical data to make predictions about the future, they run the risk of overfitting the training set and potentially generating duplicates of the observations. Recent studies [25], [26] have demonstrated that integrating autoregressive models with diffusion models [27] substantially improves their generation speed, thereby enhancing the performance of autoregressive models.

D. Normalizing Flows

Normalizing flows [28], [29] are a class of generative models designed to transform complex data distributions into simpler, more tractable forms such as Gaussian distributions. This is achieved through a series of invertible transformation layers, and by stacking such layers, normalizing flows are able to map an intricate distribution into a simpler one [30]. The prerequisite for the invertibility of a transformation layer is crucial, as it fulfills two functions: it should permit the transformation of complex data into a more manageable distribution for analysis purposes, and it must also facilitate the creation of new data instances from this simplified distribution. To optimize the model, a tractable marginal likelihood is computed, which requires each transformation layer to be capable of calculating its Jacobian determinant efficiently. Some notable works include RealNVP [31], GLOW [32], and Residual Flow [33]. Normalizing flows are noted for their capacity to learn features and quick generation process. However, the stringent requirement for transformation modules to be invertible often makes it hard to choose more flexible network structures, resulting in inferior quantitative performance in density modeling. Despite this, their ability to perform exact likelihood calculations makes them a valuable tool for various tasks, including sample generation, latent variable projection, and density value estimation [34], [35], [36].

E. Diffusion Models

Diffusion models have achieved noteworthy accomplishments within the field of generative models. It first attracted extensive interest and widespread recognition with the publication of the paper titled 'Denoising Diffusion Probabilistic Models' [27] in 2020. Similar ideas also came to the attention of the public when Score-based Generative Models [37] were proposed, bridging both the diffusion model and score-based generative model into a unified framework. There are endeavors focusing on theoretical or engineering optimization such as accelerating sampling speed, like DDIM [38] and DPM-Solver [39], or cutting down training cost, like Stable Diffusion [40]. These works enhance the practical performance of diffusion models, making them more tractable for either training or inference. Concurrent with its rapid development, the diffusion model has emerged as a prominent generative model known for its strong theoretical foundation and exceptional performance. It has been widely applied in various downstream applications like image inpainting [41], image-to-image translation [42], image composition [43], image customization [44] and prompt editing [44]. ControlNet [45] was proposed to additionally integrate various applications within a single framework. Beyond images, diffusion models also succeed in generating contents of other modalities, including videos [46] and 3D objects [47].

III. Visual Signal Coding With Generative Models

This section provides a concise review of the application of generative models to visual signal coding, focusing primarily on image and video coding.

A. Image Coding With Generative Models

In fact, the phrase "image coding with generative models" can have multiple meanings. On the one hand, probabilistic generative models provide the theoretical foundations for end-to-end learned image coding, sometimes referred to as learning-based image coding, neural network-based image coding, or neural image coding in the literature. Specifically, probabilistic generative models, such as variational autoencoders (VAEs) [15] and diffusion models [48] contribute to successful frameworks for neural image coding [49], [50]. Also, probabilistic generative models, such as autoregressive models [23] and normalizing flows [28], inspire several important improvements in coding performance [51], [52]. On the other hand, some well-established generative models, such as Generative Adversarial Networks (GANs) [1] and diffusion models, can be combined with these end-to-end learned image coding models, which have been demonstrated to provide better perceptual quality. In this section, we overview techniques in both areas, and review an important theory in the field of generative image coding: the rate-distortion-perception tradeoff.

1) Probabilistic Generative Models for Image Coding: Prevalent methods for neural image compression follow a variational autoencoder framework. Specifically, the input image $\boldsymbol{x}$ is usually mapped into its latent representations $\boldsymbol{y}$, which are then quantized into $\hat{\boldsymbol{y}}$. Since the gradient of the scalar quantization function is zero almost everywhere [53], most methods for neural image compression employ additive uniform noise to approximate quantization during training. Early works [49], [54] connect the rate-distortion objective and variational inference in this noise-relaxed case,

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{x}\sim p_{\boldsymbol{x}}} D_{\mathrm{KL}}(q(\tilde{\boldsymbol{y}}|\boldsymbol{x})|p(\tilde{\boldsymbol{y}}|\boldsymbol{x}))\\&\quad=\mathbb{E}_{\boldsymbol{x}\sim p_{\boldsymbol{x}}}\log p(\boldsymbol{x})\\&\qquad+\mathbb{E}_{\boldsymbol{x}\sim p_{\boldsymbol{x}}}\mathbb{E}_{\tilde{\boldsymbol{y}}\sim q_{\tilde{\boldsymbol{y}}|\boldsymbol{x}}}[\log q(\tilde{\boldsymbol{y}}|\boldsymbol{x})-\log p(\boldsymbol{x}|\tilde{\boldsymbol{y}})-\log p(\tilde{\boldsymbol{y}})].\end{aligned}\tag{1}$$

Since the image $\boldsymbol{x}$ is given in the task of compression, the first right-hand-side term in the above equation is a constant during optimization. The second right-hand-side term evaluates to zero as we employ additive standard uniform noise as a stand-in for quantization during training:

$$q(\tilde{\boldsymbol{y}}|\boldsymbol{x})=q(\tilde{\boldsymbol{y}}|\boldsymbol{y})=\mathcal{U}(\tilde{\boldsymbol{y}}|\boldsymbol{y}-0.5,\boldsymbol{y}+0.5)=1.\tag{2}$$

The rest two terms in Eq. (1) denote the distortion and rate, respectively. Therefore, the rate-distortion optimization

objective in lossy compression can be successfully interpreted from the view of the variational inference. Such a joint rate-distortion optimization objective is critical to achieve promising compression performance. Despite a short history, the rate-distortion performance of this variational image compression framework has been demonstrated to surpass traditional image compression standards, in terms of both the objective metrics such as RGB PSNR [55] or MS-SSIM [56], and perceptual quality [57]. In addition to noise-relaxed quantization, some alternatives, e.g. vector quantization [58], [59], and soft-to-hard annealing [60], have been proposed to replace additive uniform noise during training. However, the mainstream approaches for neural image compression still adopt additive uniform noise during training, since it is stable during training and theoretically sound as variational autoencoders. Soft-then-hard two stage quantization [53] scheme was further proposed to learn an expressive latent space softly, then closes the train-test mismatch with hard quantization.

Recently, as a new member of probabilistic generative models, diffusion models have attracted increasing attention due to their strong capability to model data distributions [61]. Theis et al. [50] proposed a promising framework that applies Denoising Diffusion Probabilistic Models (DDPMs) [27] to neural image coding, where the image information is decomposed into posterior distributions in multiple diffusion steps and compressed with relative entropy coding [62], [63]. This new framework exhibits encouraging compression performance in terms of both objective metrics and perceptual quality. Diffusion models operate by iteratively transforming an input image into a noise distribution, allowing for effective compression while preserving essential features. However, due to the inherent huge complexity of DDPMs and relative entropy coding, this diffusion-based lossy compression framework still has ample room for further improvement.

Autoregressive models for image generation were first proposed as PixelCNN [23]. Although generation and compression are fundamentally two different tasks, the concept of autoregression inspires the design of the context model, which significantly enhances the compression performance of of neural image compression models. By decoding the latent variables sequentially in raster-scan order, the context model can improve the entropy modeling of latent variables, and therefore boost the compression performance. Subsequently, the spatial autoregressive context model was proposed to be replaced by the channel autoregressive context model [64], which leverages the correlation among latent channels and is more efficient for decoding.

Another typical type of probabilistic generative models is normalizing flows. Although the applications of normalizing flows in neural image compression are not as many as the previous three categories of probabilistic generative models, normalizing flows still play an important role for image compression. For example, normalizing flows can help the neural image compression model to achieve better idempotence [65], which means that a codec can support successive image compression [66]. The ANFIC [67] and its subsequent studies [68], [69], [70] offered a comprehensive analysis of applying Normalizing Flows to lossy image and video compression. The iWave work [71] proposed a wavelet-like transform based on normalizing flows for lossy image compression with improved performance. In addition, advanced normalizing flows can also have wide applications for lossless compression, such as integer normalizing flows [72], [73]. While we focus on lossy compression in general, the entropy coding module in the lossy compression framework is basically a lossless compression module.

In summary, the aforementioned four categories of probabilistic generative models contribute to both the theoretical foundations and the technical improvements for neural image compression, fostering significant progress in this domain.

2) Generative Image Coding for Perceptual Quality: It is known that objective metrics such as PSNR have discrepancies with human perception [74]. Therefore, more and more researchers start to pay attention to the research of image coding for perceptual quality. In addition to the success of neural image compression methods based on variational autoencoders (VAE) and other probabilistic generative models, recent advances in image compression have seen promising results through the combination of generative models and compression models, especially by combining compression models with GANs and diffusion models [27]. These approaches provide different strategies to achieve better rate-distortion performance and address concerns about perceptual quality without significantly increasing complexity.

GANs consist of a generator and a discriminator network. These networks are trained adversarially. In the context of image compression, GANs have been employed to generate realistic and high-quality images, while simultaneously optimizing compression efficiency. The generator learns to produce compressed representations of images, and the discriminator evaluates the authenticity of the generated images, creating a dynamic interplay that enhances both the visual fidelity and the compression effectiveness. Agustsson et al. [75] propose a GAN-based image compression architecture operating at extremely low bitrates, which can synthesize hard-to-store details and reduce artifacts. In addition, semantic maps can be useful for efficiently synthesizing less significant areas, if they are accessible. HiFiC [57] investigates novel structures such as conditional GANs and normalization layers to preserve fidelity while generating visually pleasing results. Multi-Realism [76] designs a model with distortion-realism trade-off, allowing users to control the level of detail in reconstructed images.

Most recently, combining diffusion models and compression models has gained increasing attention. Unlike building a new lossy image compression framework with diffusion models as mentioned in Section III-A1, such a combination can achieve a good balance between complexity and perceptual quality. Yang et al. [77] propose a novel image compression framework which applies a VAE-style encoder to map images to latent variables and a diffusion model as the decoder conditioned on quantized latents. DIRAC [78] leverages a diffusion model to enhance perceptual quality by producing residuals conditioned on an initially reconstructed image, which is decoded by an image codec with minimal distortion. Furthermore, HFD [79] explores an advanced noise schedule and sampling procedure

and designs a patch-wise approach for high-resolution image reconstruction. All these works effectively showcase the potential of integrating generative models with compression models to enhance the visual quality of decoded images.

Due to the different methods used for perceptual optimization, the learned codecs exhibit different distortion types than traditional codecs. The distortions of traditional codecs generally include blocking artifacts, blurring, aliasing, etc., while learned codecs may produce smoothing, generated noise, pseudo-texture and other types of distortions, which may vary depending on the optimization method and structure.

The substantial differences in perceptual quality among different codecs therefore require more demanding criteria for the study of accurate and explainable visual quality assessment metrics, which is also key to codec optimization. To this end, the MPEG Visual Quality Assessment ad-hoc group (AG 5) has been working to maintain a dataset of Compressed Video for study of Quality Metrics (CVQM). During the 144th MPEG meeting, MPEG AG 5 issued a call [80] for learning-based video codecs for the study of quality assessment. This is because MPEG anticipates that the reconstructed videos compressed with learning-based codecs will have different types of distortions compared to those produced by the traditional block-based motion-compensated video coding. MPEG will consider inviting responses that meet the call's requirements to submit compressed bitstreams for further study and potential inclusion into the CVQM dataset.

3) The Rate-Distortion-Perception Trade-off: The theoretical foundations of lossy compression in mathematics are rooted in Shannon's seminal work on the rate-distortion theory [81], where the distortion term is usually measured by objective metrics such as PSNR. However, in recent years, it has become increasingly accepted that 'low distortion' is not a synonym for 'high perceptual quality'. In fact, optimizing one often comes at the expense of the other [82]. Following the mathematical notion of perceptual quality in [74], the perfect perceptual quality is achieved when

$$p_X = p_{\hat{X}}, \tag{3}$$

where the input image is X, the decoded image is $\hat{X}$, and p_X is the distribution of the input images. On the basis of this, the work of [82] provides a systematic study of the rate-distortion-perception trade-off. An important conclusion is later discovered by the authors in [83] and [84] who assert that "We proved that, for fixed bit rate, the cost of imposing a perfect perception constraint is exactly a doubling of the lowest achievable MSE." In other words, the PSNR of the decoded image with perfect perceptual quality is 3dB lower than the PSNR of the decoded image with the smallest distortion at the same bitrate. This discipline provides an insightful guide for neural image compression models targeting perceptual quality.

In addition, Freirich et al. [85] establish the achievable distortion perception region and provide a geometric interpretation of the optimal interpolator in Wasserstein space. Chen et al. [86] study the subtle differences between the weak- and strong-sense definitions of perceptual quality and analyze the role of randomness in encoding and decoding. While most of these works start their analyses with a toy example, the work of [87] successfully applies pre-trained unconditional generative models to real-world images and is able to achieve better perceptual quality as established in Eq. (3) by proposing to pursue the idempotence in neural image compression.

In short, the rate-distortion-perception trade-off is attracting increasing attention in both theories and applications to practical problems. We believe that research on such a trade-off will continuously contribute to better approaches for generative image coding.

B. Video Coding With Generative Models

While the field of neural image compression has been fully developed, the field of neural video coding is also experiencing tremendous development. In this section, we discuss the neural video coding frameworks from two categories, as depicted in Fig. 1: the autoencoder-based coding models and hybrid coding models. Moreover, we analyze the development of generative video coding for perceptual quality improvement.

1) Autoencoder-Based Coding Models for Video Compression: A few works regard neural video compression as an extension of the autoencoder-based image compression framework, where the coding pipeline is generally divided into two parts: transform coding and conditional entropy coding. Specifically, a 3D or 2D autoencoder is used to transform the video sequence into quantized features to encode, and then a conditional entropy coder that combines spatiotemporal information is used for entropy coding [88], [89], [90]. Given the independence of distortion introduced in the time domain, this framework effectively avoids issues like error propagation. However, the algorithm's overall complexity tends to be high. Redundancy removal primarily occurs in the entropy coding module, which does not fully capitalize on the benefits of transform coding.

2) Hybrid Coding Models for Video Compression: The current mainstream neural video compression framework still uses a hybrid coding framework that combines inter-frame motion estimation, which is similar to the traditional video coding framework. Generally, the coded motion information and the previous decoded frame are used to generate the reference frame, while the residual information between the current frame and the reference frame is encoded at a later stage.

An early-stage study [91] conducted in early 2018 introduced a block-based hybrid generative module called PixelMotionCNN to model spatiotemporal coherence; the module utilizes effectively predictive coding together with additional components of iterative analysis/synthesis to reach comparable compression results with an H.264 codec. Lu et al. [93], [98] proposed to replace every module in a traditional video compression framework with neural networks. In particular, the estimated optical flow is treated as motion information and encoded together with the residual frame by two different autoencoders. At the same time, Rippel et al. [92] also proposed a highly complete model based on similar ideas, which not only replaced the traditional coding module, but also expanded it into a more general model that compresses the generalized state. Since then, numerous studies have

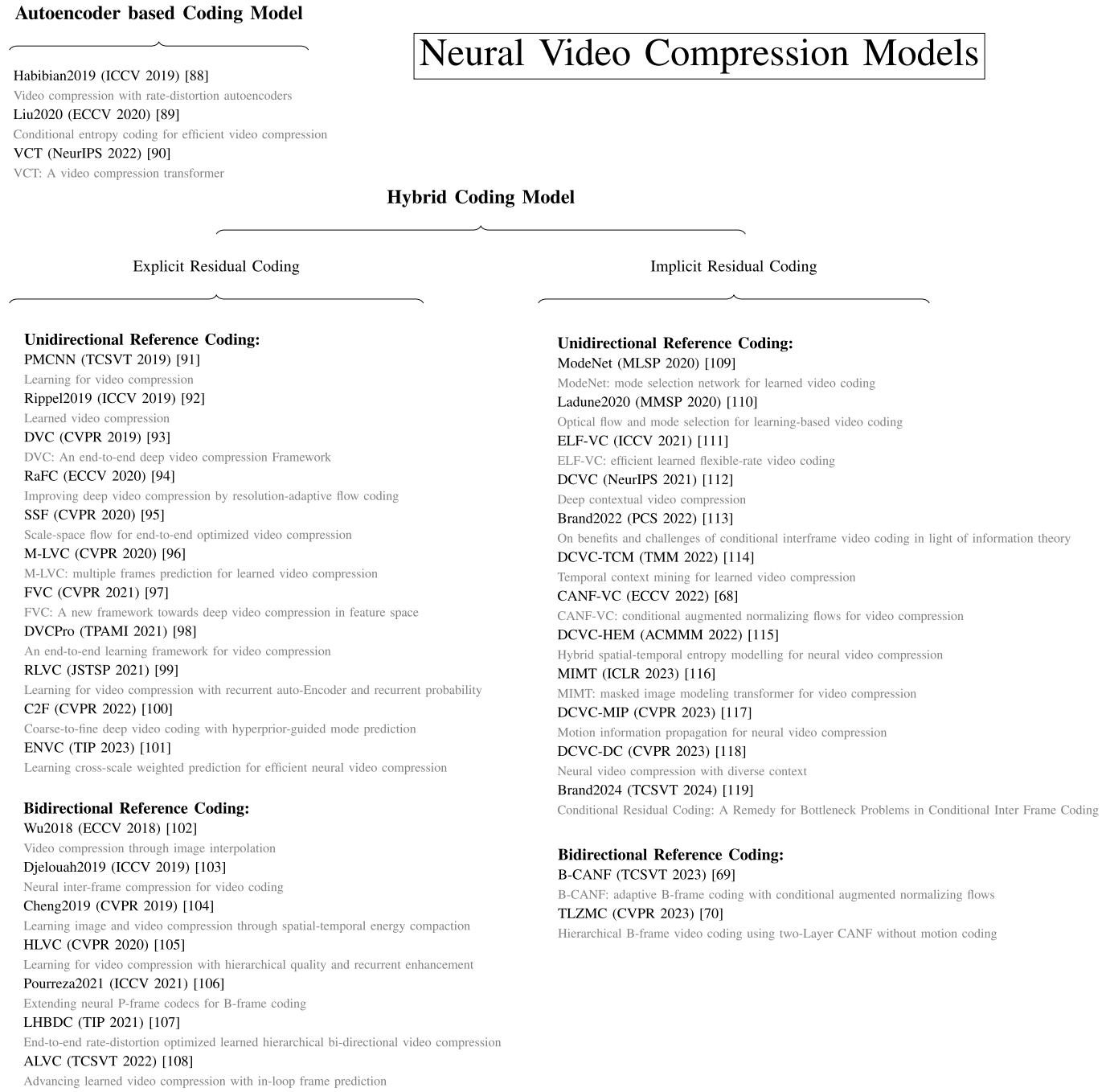


Fig. 1. Neural video compression models.

been dedicated to exploring methods for acquiring effective representation and precise modeling of motion information. Some works attempt to perform hierarchical processing in high-dimensional representation space to achieve a better rate-distortion balance [97], [100], while other works explore better motion estimation and compensation [95], [96], [101], motion compression [94] or time correlation mining [99] to improve the accuracy of representation. In addition, some works start from the reference relationship between frames and use frame interpolation models, recurrent neural networks, etc. to explore the value of bidirectional reference relationships [102], [103], [104], [105], [106], [107], [108].

Furthermore, by utilizing a hybrid coding framework in conjunction with motion estimation, a different set of studies have aimed to eliminate the need for residual frames. Instead, these studies focus on utilizing the information obtained from motion compensation results to enhance the encoding and decoding process of the frame being encoded. The former can be considered as explicit residual coding, while the latter can be considered as implicit conditional coding. Theoretically, as was demonstrated and discussed in [113], the latter has a higher rate-distortion upper bound. In order to introduce the skip mode structure, Ladune et al. [109] firstly built and tried out the concept of conditional coding in the area of

learned video compression, and developed their architecture with an optical flow network in [110]. While [111] introduced a more general design of conditional coding from an engineering perspective, [112] further summarized the core idea and designed an efficient coding model based on it. The follow-up work focuses on more efficient information transmission [114] and functional improvements such as supporting variable rates [115]. Several following works are put forward through architecture and module improvements [68], [116], [117], [118]. The later work [119] re-added the residual structure in conditional coding to solve the possible bottleneck problem. As the representative work of implicit conditional coding, the work [118] has been able to surpass the low-delay configuration of the H.266/VVC standard reference software VTM in terms of both the objective metrics of RGB PSNR and MS-SSIM.

3) Generative Video Coding for Perceptual Quality: Although most neural video compression works still strive to improve objective performance with metrics such as PSNR and MS-SSIM, a few works have drawn inspiration from image compression techniques to enhance the visual quality of compressed videos. Yang et al. [120] investigated perceptual optimized video compression with recurrent conditional GAN. Mentzer et al. [121] directly explored conditional GAN training and verified the effectiveness of this method through user studies, while another work analyzes perception loss functions for learned video compression [122]. The work [123] presents a diffusion probabilistic modeling approach for video generation, drawing inspiration from recent advances in neural video compression, which has the potential to offer valuable insights for enhancing the perceptual performance of video coding.

IV. Standardization Activities Related to Visual Signal Coding With Generative Models

This section briefly overviews the standardization activities related to visual signal coding with generative models. Specifically, the progress of image and video coding using generative models is described in the following subsections.

A. JPEG AI Standardization Activities

Traditional image coding schemes, such as JPEG [127], JPEG2000 [128] and intra coding in H.264/AVC [129], H.265/HEVC [130], AV1 [131] and H.266/VVC [132], are developed in a classical paradigm, including block partition [133], intra prediction [134], transformation [135], and entropy coding [136]. Recently, learned image compression methods [51], [54], [137] have achieved significant progress and superior improvements in rate-distortion performance, attracting lots of attention from both industry and academia. One attractive feature lies in that learned image coding schemes are able to adapt to different applications with moderate additional effort. They can achieve good compression results for machine vision and image processing tasks by changing the optimization targets [138].

JPEG AI, a learning-based image coding standard, was under development [138] at the time of writing and was expected to be finalized in 2024. JPEG AI is the first AI-based image coding standard developed jointly by the ISO/IEC and ITU-T. The scope of JPEG AI is the creation of a learned image coding standard aimed at offering a royalty-free baseline that achieves significantly better compression performance than existing image coding standards and meanwhile provides enhanced performance for image processing and computer vision tasks. JPEG AI is designed to cater to versatile real-world image applications, such as surveillance, cloud and edge storage, autonomous driving, and visual data transmission and distribution. The Working Draft (WD) and the Committee Draft (CD) of the JPEG AI core coding system were released in 2023. The international standard is expected to be published in 2024.

Analogously to many existing learning-based image coding schemes, JPEG AI adopts a VAE-based framework [51], [54]. The input image is compactly converted into its latent representation by an analysis transform network at the encoder side. The synthesis transform network recovers the input image from the latent representation at the decoder side. Moreover, the latent representation is entropy coded as a bitstream. The framework of JPEG AI is illustrated in Fig. 2, where only the modules in red color are standardized. The overall framework is optimized in an end-to-end way by minimizing the rate-distortion cost, where the distortion term can be calculated with the Mean Squared Error (MSE) or Multi-Scale Structural SIMilarity (MS-SSIM) [139] for optimized objective quality during training. Generative models such as GAN have been trialed for optimizing perceptual quality during the development of JPEG AI [125]. In particular, the GAN-based perceptual loss [140] is introduced as one of the distortion terms for training the networks, generating perceptual-oriented models that provide better visual quality in low-rate coding scenarios.

JPEG AI Common Test and Training Conditions (CTTC) [141] present guidelines on the training and performance evaluation of the JPEG AI Verification Model (VM). The training dataset includes 5000 images, and the validation dataset contains 350 images. The test dataset has 50 images. Various quality metrics, including MS-SSIM, Video Multimethod Assessment Fusion (VMAF) metric [142], Visual Information Fidelity (VIF) [143], are employed to evaluate the coding distortion of the reconstructed image, emphasizing the perceptual quality of human vision. The target bit-rates are {0.06, 0.12, 0.25, 0.50} bits-per-pixel (bpp). The coding complexity is measured in the number of multiply-accumulate operations per pixel (kMAC/pxl). Compared with the VVC intra coding [144], JPEG AI VM-4.3 achieves a 28.5% BD-Rate saving at the high operation point. At the base operation point, the averaged coding gain is 16.4% [145].

1) VAE-Based Transformation: The analysis and synthesis transform networks of JPEG AI are built with residual blocks, attention blocks [146], and activation layers for nonlinear conversion. These networks are operated with the YUV 420 color format by default, such that a color space conversion layer is provided at the beginning of the analysis transform at the encoder side. Moreover, the luma and chroma components employ separated analysis and synthesis transform networks

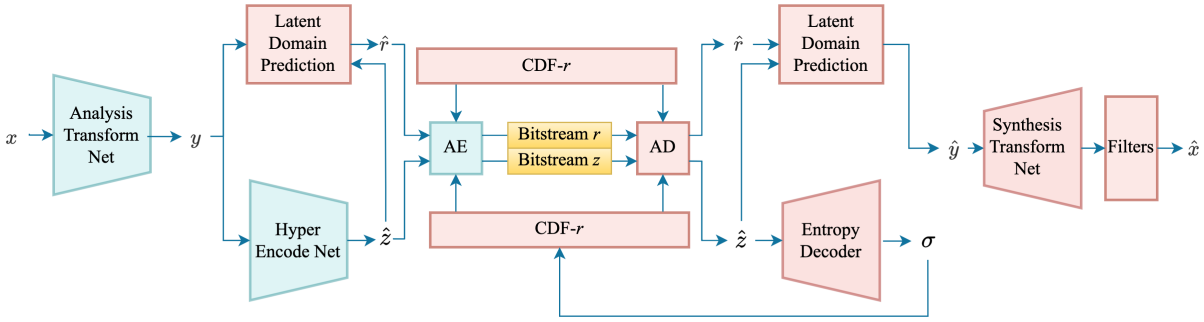


Fig. 2. VAE-based JPEG-AI framework. x and $\hat{x}$ denote the original input image and the reconstructed images, respectively. The red modules are standardized in JPEG-AI. The blue-green modules are the encoder side operations [124].

to reduce peak memory usage. The network is deeper and heavier for the primary component (luma) to better compress and reconstruct the texture details. As such, given the input image, the analysis transform cooperating with latent domain prediction tools maps the image into its latent representations, which include the residual, prediction and entropy information. The synthesis transform generates the visual signals based on the latent representations parsed from the bitstream. The hyper-prior networks, including the hyper encoder network, hyper decoder network and hyper-scale decoder network, are also built on top of the VAE, where the distribution parameters are embedded as the hyper-prior term for high-efficiency entropy coding.

During the standardization of JPEG AI, two operation points, namely, the base operation point and the high operation point, are developed to cater to different application scenarios. The layer design such as the convolution kernel size and the upscaling strategy of these two operation points are different. The decoding complexities of the base operation point and the high operation point are around 20 kMAC/pxl and 200 kMAC/pxl, respectively [145]. In particular, the base operation point is with minimal networks for encoding and decoding, providing a lightweight decoder suitable for the deployment on mobile devices. The coding tools such as the residual and variance scale (RVS), latent scale before synthesis, and enhancement filters are all disabled in the configuration of the base operation point. At the high operation point, attention models, including the Transformer-based Attention Module (TAM) and the Convolutions-based Attention Block (CAB), are enabled in the synthesis transform network to enhance the generative capability of the decoder [124]. The high operation point enables all the coding tools for enhanced compression performance.

2) AR-Based Context Modeling and Entropy Coding: A decoupled architecture [125], [126] is designed for entropy coding to decouple the decoding dependencies between entropy decoding and latent sample reconstruction, as illustrated in Fig. 3. Unlike the previous design, where the arithmetic decoder is interleaved with the generation of the entropy parameters, the entropy decoding process in the decoupled architecture is independent from the latent sample reconstruction, leading to significant reductions in decoding time. This design philosophy has been adopted by traditional video coding standards. To be more specific, a hyper decoder

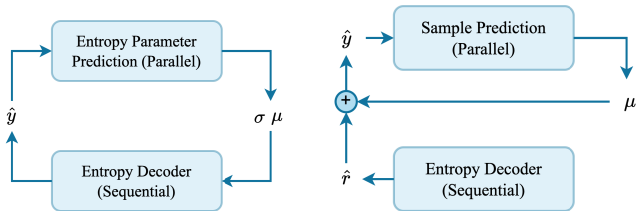


Fig. 3. Illustration of the sequential architecture and the decoupled architecture; left: sequential architecture; right: decoupled architecture [125], [126].

and a hyper-scale decoder are involved in the decoupled architecture. The prediction of latent samples is reconstructed from the hyper decoder, and the Gaussian variance is recovered from the hyper-scale decoder. After the quantized residual samples are obtained from the bitstream, the reconstruction process is invoked. In this way, the latent reconstruction and the entropy decoding are separated; consequently, the entropy decoding will not be suspended by the latent reconstruction.

Autoregressive models exhibit prominent context modeling capability, which has been widely employed in the learned image compression. Following the raster-scan sequential processing order, the reconstruction of the current sample relies on the previous neighboring reconstructed samples. As such, the main issue of the autoregressive model lies in the strict sequential processing during context modeling, which hinders its deployment in real-world applications. To facilitate parallelization and improve GPU utilization efficiency, the wavefront parallel strategy is supported in JPEG AI context modeling [125], [126], with which the latent can be constructed in the row-wise concurrent processing order, as shown in Fig. 4. The context network involves neighboring reconstructed samples as input and yields multiple outputs simultaneously, leading to much reduced decoding time.

Even though the wavefront decoding strategy can enhance the parallelization of the context modeling and latent prediction, there is still room for further improvements to the prediction speed. A Multi-stage Context Modeling (MCM) method [147] is adopted in JPEG AI, which enhances parallelization to save decoding time and maintain coding performance. More specifically, the MCM is built on top of the decoupled architecture, which is used as the replacement of the context model when predicting the mean of the latent representation, so as to further reduce the decoding complexity of the entropy coding module. Instead of sequentially modeling the contexts, the tensor of the latent representation is

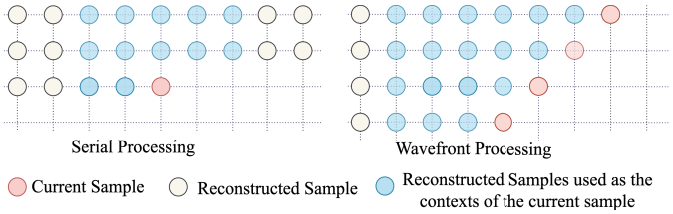


Fig. 4. Illustration of the serial processing based on the raster-scan order and wavefront processing [125], [126].

divided into 8 groups through down-shuffle, and sub-groups are concatenated in the channel dimension. The context modeling process can be regarded as the implicit prediction of the element in the latent representation with the reconstructed group elements. MCM achieves 97% speedup during the latent sample prediction in the entropy decoder, with only a 2.1% BD-Rate loss.

B. Exploration of Neural Network-Based Video Coding

Dating back to June 2020, the Joint Video Exploration Team (JVET) of ITU-T VCEG and ISO/IEC MPEG established an "ad hoc group" (AhG) on NNVC at the 130th MPEG meeting and 19th JVET meeting [148]. Since then, two categories of NNVC have been extensively studied within this AhG. In the first category, the neural network-based (NN-based) modules are embedded in the traditional video coding framework to improve coding performance. In particular, the NN-based modules are used to replace the classical coding modules such as intra prediction [149], [150], [151], [152], inter prediction [153], [154], in-loop filter [155], [156], [157], [158], [159], post filter [160] and resampler [161], [162]. The overall framework is optimized with rate-distortion criteria following the traditional coding philosophy. In the second category, the coding paradigm is achieved with the full neural network. More specifically, predictive coding is explored, which employs optical flows to realize the inter prediction. Then, residual signals are derived and entropy coded with an autoencoder. Conditional coding is also investigated in the full neural network based video coding, wherein the prediction is achieved by deriving the latent representation with the autoencoder [163], [164]. It is noted that the proposed end-to-end learned coding method in [163] is only applied to inter frames. For coding intra frames (i.e., I frame), the traditional video coding standards, such as H.266/VVC intra coding or BPG compression, are applied. Meanwhile, a common test model, also knowned as Neural Network-based Video Coding (NNVC), was initially produced with two NN-based in-loop filtering tools in July 2022 and maintained for the exploration experiments on NN-based technologies. NNVC evolves with the investigation activities of each meeting cycle. As of Jan. 2024, NNVC-7.1 comprises two main NN-based modules including the NN-based intra prediction and NN-based in-loop filtering, with the goal of enhancing the coding performance of the traditional tools within the current H.266/VVC standard. To be more specific, with the NN-based intra prediction, fully connected neural networks are utilized to establish a nonlinear mapping between neighboring reference samples and samples in the current block [150], [165]. Moreover, the NN-based module additionally produces auxiliary outputs that assist in constructing the Most Probable Mode (MPM) list and selecting transform kernels for subsequent processes. Regarding the NN-based in-loop filtering, a convolutional neural network-based in-loop filter is utilized to enhance the reconstruction quality and recovering capability [157], [166], [167], [168], [169]. The NN-based filter undergoes iterative training to tackle the problem of excessive filtering, as outlined in [169]. Enhancing performance involves additional considerations such as leveraging coded information, selecting parameters, adapting inference granularity, scaling residuals, incorporating temporal filtering, integrating deblocking filtering, aligning with Rate-Distortion Optimization (RDO). In addition, the deep filter supports two trade-off points in terms of compression complexity and efficiency, namely the low operation point (LOP) at 17kMAC/pixel and high operation point (HOP) at 477 kMAC/pixel. The coding performance of the NN-based filter and intra prediction in NNVC-7.1 as compared to the H.266/VVC reference software VTM-11.0 is summarized in Table I [170]. From this comparison, it is evident that significant BD-rate gains of up to -13.59% and -12.55% are observed for the Y component under Random Access (RA) and All Intra (AI) configurations, respectively. This underscores the considerable potential of neural network-based coding tools in advancing video compression performance.

TABLE I
CODING PERFORMANCE (BD-RATE) OF NNVC-7.1 OVER H.266/VVC REFERENCE SOFTWARE VTM-11.0 [170]

Methods	AI			RA		
	Y	U	V	Y	U	V
LOP+NN-Based Intra	-8.13%	-13.28%	-13.42%	-6.89%	-13.17%	-12.15%
HOP+NN-Based Intra	-12.55%	-11.37%	-13.05%	-13.59%	-12.47%	-14.18%

C. Exploration of Generative Face Video Coding

Different from early model-based coding (MBC) techniques [171], [172], [173], [174], generative face video coding (GFVC) schemes [175], [176], [177], [178], [179], [180], [181], [182] exploit the excellent generative ability of deep generative models [1], [183] to improve the face reconstruction quality and realize ultra-low bitrate face video communications. Specifically, the encoder employs the traditional image/video codec to compress the key-reference frames of a face video, and encodes the subsequent inter frames into compact transmitted symbols (*e.g.,* landmarks/keypoints, compact feature and facial semantics). Besides, the decoder feds these decoded key-reference frames and compact facial representations into the deep generative model to learn the temporal evolution and reconstruct these face frames. The typical framework of GFVC is depicted in Fig. 5.

Inspired by such promising rate-distortion performance, some GFVC proposals have been submitted to JVET, where they explored whether GFVC's compact symbols could be inserted into an H.266/VVC bitstream as Supplemental Enhancement Information (SEI) messages. More specifically, the generative face video (GFV) SEI message [184], [185] allows VVC-coded pictures to be utilized as the base

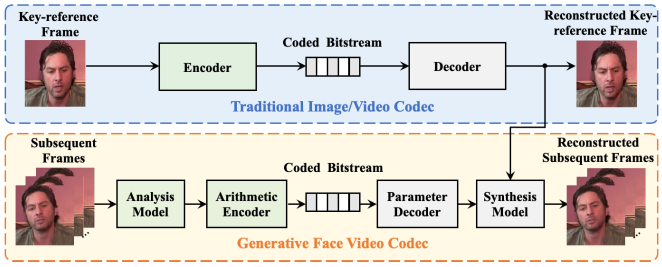


Fig. 5. Framework of Generative Face Video Coding.

(key-reference) pictures and incurs only very little overhead in the compressed bitstream to signal compact facial information with different representation types, including compact temporal features, 2D keypoints, 3D keypoints, facial semantics and others. In addition, several proposals [186], [187], [188] aim to enable a more common GFV SEI syntax and specify the decoder interface with the generative neural network.

Furthermore, Ye et al. [189] made the related technical requirements regarding the exploration and potential standardization of ultra-low bitrate 2D generative face video coding methods. As such, JVET experts decided to establish a new ad hoc group to conduct GFVC investigations on software implementation, experiment coordination, interoperability studies, and other related aspects. In particular, a unified software package [190] with various GFVC methods has been proposed to allow coding to be performed using the VVC Main 10 profile in Jan. 2024. Besides, the common test conditions and software reference configurations [191] have also been specified for GFVC experiments. In addition, the interoperability study in feature translation [192] and lightweight model [193] on different GFVC systems are further investigated to allow more flexible GFVC applications within acceptable performance losses.

In short, the current GFVC standardization activities are mainly concentrated on the design of the SEI message with rich facial representations such that it warrants the service of ultra-low rate communications, user-specified animation/filtering and metaverse-related functionalities. Certainly, there exist issues and challenges for GFVC's standardization and deployment, e.g. unstable generation quality, high decoder complexity, inadequate model interpretability and inappropriate evaluation measures.

V. Visual Signal Processing With Generative Models

Apart from visual signal coding, another major application venue of generative models is visual signal processing. Various generative models have been applied and adapted to different image and video processing tasks, including restoration, synthesis, editing, and interpolation. This section provides a brief overview of the important works on these research topics, highlighting the important role of generative models for these applications. It is noted that there is another important research area focusing on the generation of visual signals from texts. Due to the limited length of this paper, here we solely review the approaches that take visual signals as input.

A. Generative Visual Signal Restoration

Image and video restoration is a process to recover the high-quality version of a visual signal that is associated with different perceptual artifacts. These artifacts can be generated in different production stages, including content capture, transmission, and display, which could potentially degrade the perceptual quality of visual signals and reduce the effectiveness of high-level computer vision algorithms (e.g., detection and classification) [194]. Various restoration methods are commonly applied at different steps in the production pipeline. According to the nature of the artifacts, restoration methods can be classified as denoising (to remove camera and production noises), deblurring (to reduce focal or motion blurs), dehazing/deraining (to alleviate the haze and raining effect), super-resolution (for spatial resolution upsampling) and compression enhancement (to remove compression artifacts) [195]. As the review of visual signal coding with generative models has already been conducted in Section III, here we solely focus on other restoration tasks, providing a concise summary of some key works based on generative models, which are proposed for visual signal restoration. For a more comprehensive overview of the literature on this topic, the reader is referred to references including [195], [196], [197], [198].

1) VAE-Based Restoration: Early attempts at generative restoration include [15], [199], which employ vanilla VAEs to perform denoising. Their performance has been further improved by Denoising Autoencoders (DAE) [200], which train VAEs with noise injected into their stochastic hidden layer. Moreover, the DAE model has been enhanced through the combination with a more advanced training methodology, resulting in Denoising Adversarial Autoencoders (AAEs) [201], and by incorporating explicit models of the image noise distribution in the decoder, with DivNoising [202] as a notable work. VAE-based approaches have also been applied to deblurring, with examples including [203], [204], where the networks employed consist of an autoencoder to learn the image prior and an adversarial network to discriminate blurred and clean images or their features. Moreover, VAEs have also contributed to the task of dehazing/deraining. The variational image detraining (VID) [205] method uses a conditional variational encoder (CVAE) to perform probabilistic inference that increases the diversity of prediction. pWAE (pixel-wise Wasserstein autoencoder) [206] introduces 2D latent tensors to the Wasserstein autoencoder to allow pixel-wise matching, which is reported to offer better dehazing performance compared to conventional autoencoders. Finally, VAEs have also advanced the development of superresolution. Important works include SR-VAE (Image Super-resolution via Variational AutoEncoders) [207] and VDVAE-SR (Very Deep Variational Autoencoder Super-resolution) [208]. SR-VAE learns the conditional distribution of high resolution images providing their low resolution counterparts, which can generate super-resolution results with photorealistic visual quality and relatively low distortion [207]. VDVAE-SR adapts a very deep VAE model to single-image super-resolution and employs a low resolution (LR) encoder to learn the image prior. This has been demonstrated to provide competitive

super-resolution performance compared to the SotA at the time [208].

2) GAN-Based Restoration: As one of the primary types of generative models, GANs have been widely used for visual signal restoration. Typically, these GAN-based approaches enable the generation of photorealistic details rather than simply minimizing the distortion between the output and the training targets. For the task of denoising, a number of GAN-based methods have been proposed in the context of natural and medical image denoising. For example, a GAN was trained in [209] to learn the noise distribution within noisy input images, based on which noise samples are generated from clean images and used to train a deep CNN for denoising. For this task, more advanced GAN architectures are also employed, with examples based on CycleGAN [210], StyleGAN [211], and Wasserstein GAN [212]. In the research field of dehazing/deraining, researchers not only utilized existing GAN models for clean image and video recovery [213], [214], but also developed customized GAN-based approaches for this purpose. A notable approach is FD-GAN, which employs a fusion discriminator to learn additional priors from frequency information - this allows the generation of more photorealistic dehazed images. Similarly, DW-GAN [215] was developed by combining a GAN with a discrete wavelet transform to obtain excellent dehazing performance for images with a nonhomogeneous haze effect. For superresolution, a large number of methods have been proposed based on existing GAN models such as standard GANs [5], [216], conditional GANs [217], [218], patch GANs [219], [220], relativistic average GANs [221], [222], [223] and Wasserstein GANs [224]. One of the earliest but influential works is SRGAN [5], which is arguably the first attempt focusing on perceptually inspired super-resolution. Different GAN variants have also been designed specifically for this task, with recent examples such as content-aware local GAN (CAL-GAN) [225] and Generative and Controllable Face Super Resolution (GCFSR) [226].

3) Restoration Based on Diffusion Models: In the past three years, being one of the most popular research topics in machine learning and computer vision [227], diffusion models have now been actively exploited in the context of image and video restoration. Although it is at a very early stage, this type of approach shows the promise in competing with classic CNN-based restoration methods and those based on other generative models. For example, denoising diffusion probabilistic models (DDPMs) have been employed for single- and multiple-weather image restoration (e.g., desnowing, deraining, and dehazing) in [228]. DDPMs have also inspired super-resolution approaches such as SR3 (Super-Resolution via Repeated Refinement) [229] and SRDiff [230]. Moreover, a new Denoising Diffusion Restoration Model (DDRM) has been proposed in [231] for multiple restoration tasks, including debluring, super-resolution, and inpainting. While various diffusion models have been used and developed for the restoration task, they can also be combined with other deep learning techniques to achieve improved restoration performance. One of the notable works in this category is the Implicit Diffusion Model (IDM) [232], which integrates the implicit neural representation and diffusion models in the same framework - this allows the developed super-resolution model to perform continuous-resolution requirement. Despite the promising results generated by various diffusion models for the restoration task, its low computational efficiency has also been observed and considered a common drawback. Recently, efforts have been made to design light-weight diffusion-based restoration models, including Spectral Diffusion [123] and DiffIR [233].

B. Generative Visual Signal Synthesis and Editing

As another important research area in visual signal processing, image and video synthesis and editing focus on generating photorealistic content, or editing existing images or videos with a new style, background, or foreground [234]. In recent years, advances in generative models have made significant contributions to this research field. According to the input references, these synthesis and editing methods can be classified as visual guidance, audio guidance, or text guidance. Due to the limited space in this survey paper, here we mainly focus on the review of generative synthesis and editing approaches based on visual guidance. For a more detailed overview on this topic, the readers are referred to [234]. Among all generative synthesis and editing methods, an important approach is Pix2PixHD [235], which generates high-resolution synthetic images using conditional GANs from semantic label maps. Another influential work is SPADE [236], in which semantic image synthesis is achieved through spatially-adaptive normalization using a VAE. This method has been reported to generate better synthesis results compared to Pix2PixHD [235]. More recently, diffusion models have also been used for this task, with the latest examples including SDM [237] and CycleDiffusion [238]. The former exploits the use of DDPM for semantic image synthesis, while CycleDiffusion investigates the stochastic diffusion probabilistic models in the latent space, and shows its effectiveness for various image editing tasks. Specific models have also been designed for image-to-image translation, including those based on GANs [239], [240], autoregressive models [241], VAEs [242], [243] and diffusion models [244], [245].

C. Generative Video Frame Interpolation

Similar to super-resolution, which increases spatial resolution, video frame interpolation (VFI) is the technique for generating content with higher frame rates through synthetically creating intermediate frames between existing consecutive video frames. Non-generative leaning-based VFI methods are typically classified as kernel-based [246], [247] or flow-based [248], [249], based on different network structures and motion models, respectively. Recently, generative VFI methods have also been developed to obtain interpolated content with high-fidelity perceptual quality. GANs are widely adopted in this research field, typically with an adversarial network used to enhance the generator to produce results with better visual quality. The standard GAN architecture has been employed in [250] for VFI; A multiscale GAN structure was developed in [251] to achieve improved visual quality

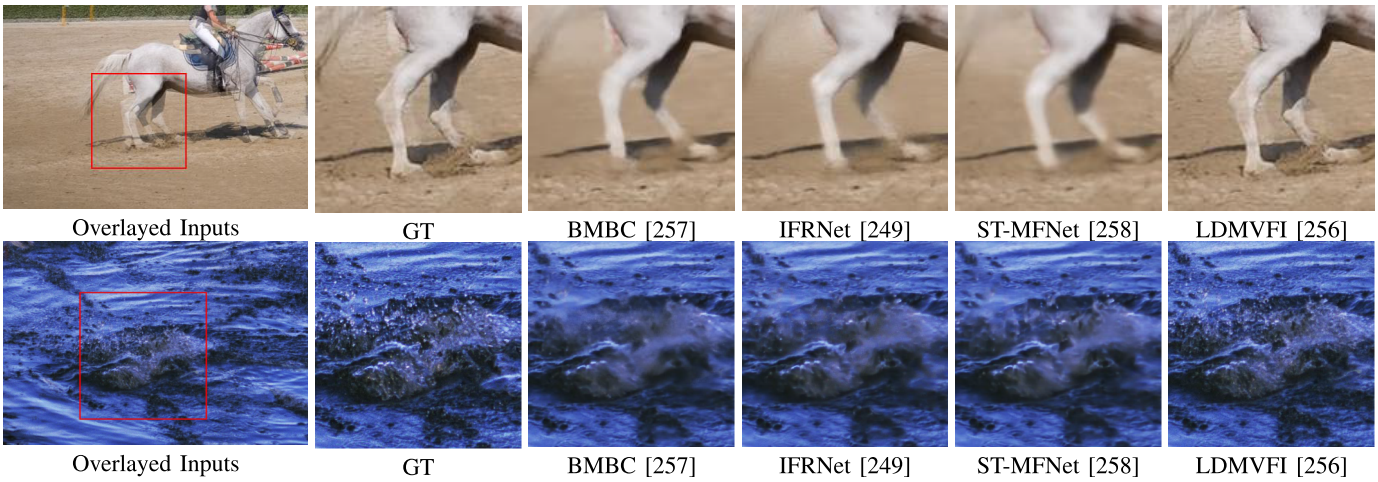


Fig. 6. Visual comparison of the video frame interpolation results generated by SotA generative and non-generative models.

and faster interpolation speed; Two GANs are concatenated in [252] to learn spatial and temporal (motion) information separately. Moreover, Frame-GAN was developed based on Wasserstein GANs with gradient penalty [253] and a modified generator loss. Recently, the research of VFI has been boosted by the advances in diffusion models. One of the first diffusion-based VFI approaches is the Masked Conditional Video Diffusion (MCVD) [254], which is based on a probabilistic conditional score-based denoising diffusion model. Its performance has been further outperformed by MV-Diffusion (Motion-aware Video Diffusion) [255], in which long-term and short-term motion trajectories have been learned using DDPMs with a motion trend attention model. More recently, latent diffusion models (LDM) have been adapted to VFI, and one of the resulting approaches, LDMVFI [256], achieves superior interpolation performance, in particular for video content with large motions and dynamic scenes. Fig. 6 shows the visual comparison results between LDMVFI and four other well-performing, non-generative VFI methods, IFRNet [249], BMBC [257] and ST-MFNet [258]. It has been observed that LDMVFI can reconstruct sharp edges and textural details, which are similar to those in the ground truth content, while other methods tend to result in interpolation results with blurring and structural artifacts. This demonstrates the effectiveness of diffusion models when employed for VFI.

VI. Generative Quality Assessment

In this section, we summarize the related work on generative model-based quality assessment and the quality metrics developed for assessing generative models.

A. Generative Model-Based Quality Assessment

Visual quality assessment is one important research topic in visual signal processing. It is typically used for evaluating and comparing the performance of different processing methods. In the context of deep learning, quality metrics can be also employed in the training process as a loss function, to optimize the model generalization. In the current literature, generative models have also been used for image and video quality assessment. For example, a no-reference image quality metric has been proposed in [259] based on a GAN, which generates training samples to tackle a common issue with deep learning based quality assessment, the lack of reliable training content. A further work [260] focuses on using a GAN to predict the primary content of a distorted image, with internal generative mechanism inspired constraints. Moreover, alternative GAN architectures have been exploited by researchers in this research field, such as [261] using Wasserstein GANs to achieve opinion-unaware image quality assessment. Recently, diffusion models (DDPM in this case) have also been integrated in the image quality assessment framework for facial content [262], to generate perturbations and estimate their influence on perceptual quality.

B. Quality Assessment for Generative Models

It is noted that generative models produce visual content which tends to have different characteristics compared to that generated by non-generative learning based and conventional processing approaches. This challenges the current common practices used for algorithm evaluation. It leads to another research area, quality assessment for generative models. To this end, researchers have developed various quality databases that contain content produced by generative models. This provides a ground-truth database for benchmarking existing quality metrics and developing new assessment methods. LCIQA [263] is one of such works, which generates test content employing various learned video codecs based on commonly used CNNs and GANs. The resulting database was then used to evaluate several widely used full reference and no reference quality metrics. Another study [264] investigates the performance of heuristic metrics such as the Inception Score (IS) [265] and the Fr'echet Inception Distance (FID) [266] for generative models in the context of image generation. The results show that although these metrics offer a relatively good correlation to several f-divergences, their ranking ability is limited when generative model performance is close. To further improve the quality prediction accuracy for generative models, enhancement methods have been proposed, including compound FID (CFID) [267], GAN-IQA [268] and DR-IQA [269]. A recent work in this domain proposed a

lightweight generalizable framework to evaluate generative models [270]. The new metrics developed in this work demonstrate improved quality prediction performance compared to existing evaluation methods, such as FID [266].

VII. Optimization of Visual Signal Coding and Processing With Generative Models

Though generative models have shown promising results in visual signal coding and processing, their implementation requires care and optimization. First, due to the use of neural networks, the complexity is often too high to meet real-time requirements. To solve this problem, there have been several fast optimization techniques at the algorithmic and architectural levels. Second, issues such as the model robustness and variable bitrate also deserve investigation.

A. Fast Optimization for Learned Image Compression

In this section, we introduce various optimization techniques for learned image compression models that adopt the factorized prior and/or hyperprior.

1) Algorithmic Optimization: Network quantization plays an important role in algorithmic optimization. As shown in [271], compared with the 32-bit floating-point arithmetic, the 8-bit fixed-point arithmetic reduces the energy consumption for additions and multiplications by 30x and 19x, respectively. Therefore, network quantization is crucial for fast and low-complexity implementations. Different from other generative applications, learned codecs require entropy coding, which demands the bit-exact accuracy to ensure interoperability across platforms. As a result, network quantization is also essential to bit-exact computation. Several quantization methods have been proposed in the literature [272], [273], [274], [275], [276], [277].

As reported in [278], there are two major approaches to network quantization: (1) the quantization-aware training (QAT) and (2) the post training quantization (PTQ). For the QAT scheme, a pre-trained floating-point network is quantized and fine tuned in the presence of quantization. The fine tuning may be carried out with respect to the network weights and/or the additional parameters (e.g. quantization bit depth) for quantization. For the PTQ scheme, the pre-trained floating-point network is directly quantized. The key procedure is to use the calibration dataset to find the optimal clipping range and quantization schemes (e.g. linear or non-linear quantization with or without zero offset).

For learned image compression, [279] exploits the Vitis AI Quantizer to generate an 8-bit quantized model. Both PTQ and QAT are supported in the Vitis AI Quantizer. After the quantization, the coding performance loss is tolerable in terms of the bpp overhead and PSNR loss. Reference [273] proposes a channel-wise QAT scheme for the weight quantization. A heuristic fine-tuning scheme starting from the synthesis transform is developed. In the case of the 8-bit quantization, the coding loss is negligible compared with the 32-bit anchor. In order to further reduce the coding loss resulting from the activation quantization, a PTQ method based on the channel splitting is proposed in [280]. Specifically, some channels with large magnitudes are equivalently split into multiple channels to reduce quantization errors while a few channels are pruned to maintain the overall network complexity. As compared with the previous work [273], it achieves a BD-rate saving of up to 4.74%. Another innovative work [281] proves that the well-used mean square error reduction is not an optimal criterion to decide the quantization parameters. Alternatively, they propose a rate-distortion optimized PTQ (RDO-PTQ), which uses the rate-distortion cost as the criterion for PTQ. Compared with [280] and [281] performs better on MSE-optimized models.

In addition to network quantization, network pruning, another type of network compression, has also been applied to learned codecs. Reference [282] aims at pruning the hyper path. Based on ResRep [283], a Lasso penalty is added in the loss function to adapt the number of pruned channels. Results show that at least 22.6% of the network parameters are saved with a negligible coding loss. Reference [284] proposes an asymmetric framework composed of a heavy encoder and a lightweight decoder. In addition, the unstructured element-wise pruning and structured channel-wise pruning methods have been trialed. Interestingly, in the case of channel-wise pruning, removing the channels with larger l_1 norms is found to be more effective than removing the ones with smaller l_1 norms.

There have been some other low-complexity algorithms. To reduce the decoding complexity, [285] realizes a real-time framework by mask decay. They utilize knowledge distillation techniques to transform the parameters from large models to small models. By doing so, the coding performance is improved by more than 30% for smaller models. In addition, the residual representation learning is proposed to implement a variable-rate encoder. Reference [286] utilizes independent separable downsampling and upsampling components to reduce the network burden. Besides, similar to [284], an asymmetric architecture is proposed to boost the decoding speed. When implemented on Intel Core i7-9700K@3.6GHz, it reaches a decoding throughput of 37.5 FPS for small models.

In addition, some hardware-oriented algorithmic optimizations have also been proposed. Reference [287] assumes that activations (feature maps) dominate the data transfer between the on-chip and off-chip memory. To reduce the bandwidth for transferring the activations, they propose a differentiable pipeline to include the required bandwidth in the classical rate-distortion loss function for training. Compared with using only the rate-distortion cost as the loss function, this modified training objective incurs little loss in coding performance. Reference [288] utilizes lookup tables to construct the hyper decoder. Compared with inferencing neural networks, both the model size and runtime are much reduced.

2) Architectural Optimization: This section further introduces some FPGA implementations. Reference [289] utilizes AMD/Xilinx Zynq UltraScale+ MPSoC ZCU104 fabricated with 16nm technology. The PL chip is XCZU7EV-2FFVC1156, which owns the resource of 504 kilo logic cell, 38 Mb memory and 1728 DSP. Reference [289] uses DPU as the hardware accelerator, and the working frequency is 350 MHz. It achieves 3.90 FPS for 720P, 1.68 FPS for

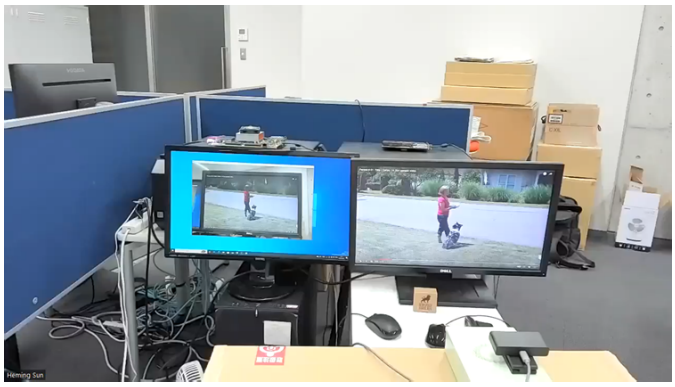


Fig. 7. A demo system for learned image compression. Encoding is performed on an FPGA board KU115, while decoding is performed on an FPGA board VCU118. The links of demo videos are given in the footnote.

1080P and 0.42 FPS for 4K resolutions, respectively. Though the throughput is rather low, as one of very early FPGA-based learned codec frameworks, [289] represents a starting point and offers useful insights into future research directions. It is a complete system, including video capturing, encoding, decoding and display. The model of [289] is mainly based on a block-wise coding framework [290].

Reference [291] also utilizes two UltraScale+ MPSoC evaluation boards ZCU102 and ZCU104 working at 200 MHz to implement a factorized model [49]. The authors implement the hardware accelerator by Verilog HDL. When dealing with 256 × 256 images, it requires 15.87 ms and 14.51 ms for encoding and decoding, respectively. Note that to reduce the hardware complexity, [291] also proposes a piece-wise linear approximation of the generalized divisible normalization (GDN) operation and its inverse operation. L eLe [292] proposes an FPGA architecture with a fine-grained pipeline to implement the hyperprior model in [293]. Different from using DPU, which is a generic architecture, the proposed pipeline architecture is more flexible for the neural layers with various CTC ratios. When implemented with an AMD/Xilinx Virtex UltraScale FPGA VCU118 development board, it achieves 40.69 FPS and 35.77 FPS for encoding and decoding 720P videos, respectively. For 1080P videos, it achieves 19.15 FPS and 16.83 FPS for encoding and decoding, respectively. Reference [294] gives a CPU-FPGA system where entropy coding is performed at the CPU side and neural computing is performed at the FPGA side. A system-level pipeline is required to process the tasks on CPU and FPGA in a parallel manner. A demo system is given in Fig. 7 where the encoding is performed on an FPGA acceleration board KU115 and the decoding is performed on VCU118. Some demo videos can be found here.[1,2]

For the above three FPGA codec systems, the power efficiency are around 29 GOPS/W, 47 GOPS/W and 21 GOPS/W for [289], [292], and [291], respectively. If those GOPS/W satisfy the required performance of learned codecs, then those codecs can run smoothly on the devices. For example, assuming one watt power supply for the circuit performing the codec operations, we can only compute about 40 giga operations per second.

However, the above results are from FPGA implementations. As compared to FPGA implementations, ASIC implementations are expected to have much higher power efficiency. However, up to now, there has been no ASIC implementation for learned image compression. One potential reason is that traditional codecs have specific components such as intra/inter prediction and DCT, so that we are able to develop corresponding ASIC chips such as [295], [296], and [297]. However, learned codecs are mainly based on neural networks, the computation of which can be executed efficiently on neural processing units (NPU). An edge device equipped with an NPU chip capable of delivering 20 TOPS/W may meet the computation requirements of some recent learned image codecs such as [298], which consumes about hundreds of GOP for one Kodak image.

B. Fast Optimization for Learned Video Compression

For video compression, since the inter frame is taken into account, the complexity becomes even higher. As a result, fast optimization techniques become even more desirable.

1) Algorithmic Optimization: Reference [299] presents a real-time design, reaching 720P@25FPS decoding on GeForce RTX 2080. To avoid the cross-platform interoperability issue, the coordinates of the transboundary quantization positions are included in the bitstream. Besides, several lightweight methods such as model pruning have been adopted to reduce the decoding complexity. As a result, decoding an I-frame takes 37.1 ms and decoding a P-frame takes 39.9 ms. With a Group of Pictures (GOP) of 12 frames, the average time per frame is 39.7 ms, which translates into 28.1 FPS. Reference [300] proposes a novel model-agnostic pruning scheme based on gradient decay and layer-wise distillation. The effectiveness has been evaluated on various learned video codecs: FVC, DCVC and DCVC-HEM. As a result, 2× speed-up with less than 0.3 dB BD-PSNR loss is achieved.

2) Architectural Optimization: As an extension of [279] and [289] gives an FPGA implementation for learned video compression with P-frame. By using the same deployment methods and evaluation board, a P-frame framework is mapped onto FPGA. When tested on JVET Class B and Class C datasets, it achieves better coding performance than x264-veryfast.

Reference [301] gives an ASIC design for learned video compression. Based on a residual coding framework, it features a CNN-Transformer neural network to enlarge the receptive field. Moreover, it adopts the Winograd algorithm to implement fast convolution and deconvolution. Notably, they develop a reconfigurable processing unit for the proposed fast algorithm. In addition, a dedicated data flow is presented to minimize the off-chip memory access. When synthesized by Synopsys Design Complier with TSMC's 28nm technology, it is able to operate at 400 MHz. With a throughput of 3525 GOPS, the real-time decoding of a 1080P video at 25 FPS is made possible. Compared with CPU and GPU implementations, this design has significantly higher energy efficiency in terms of GOPS per Watt.

[1]https://youtu.be/-unSbqsUS8Y
[2]https://youtu.be/Y4QO2h0LEDQ

3) System Optimization: Mobilecodec [302] is the first-ever real-time inter-frame learned video decoder. When tested on Snapdragon 8 chip, it achieves a decoding throughput of >30 FPS for 720P videos. Similar to traditional video compression, video frames are processed in the unit of GOP. For coding intra frames, a typical VAE-based model with the hyperprior is adopted. For coding inter frames, they follow the residual coding framework [93], which is composed of the motion network and the residual network. To reduce the complexity, a flow-agnostic motion compensation network with only convolutional operations is proposed.

To run with the fixed-point arithmetic, [302] utilizes QAT to fine tune the 8-bit quantized network. Based on the dynamic range of the computation, the channel-wise quantization is adopted for both weights and activations. Regarding the computational complexity, the I-frame decoding costs 130.9 kMAC/Pixel, and the P-frame decoding consumes 257.1 kMAC/Pixel.

As an enhanced version of [302] and [303] realizes faster throughput and better coding performance. In detail, it reduces the decoding MAC by 10 × and saves 48% BD-rate compared with [302]. The same as [302] and [303] is also based on the residual coding framework. Different from [302], a block-based warping scheme is proposed for the P-frame coding. Regarding the network quantization, the symmetric channel-wise quantization is adopted for weights, whereas the asymmetric layer-wise quantization is adopted for activations. Here, the asymmetricity represents the use of a zero offset. Reference [303] also provides a system-level pipeline for the tasks on CPU, GPU, NPU and warping kernel. When tested on HEVC-B dataset, it is able to decode the videos at 38.9 FPS. Regarding the coding performance, there is still a gap between the 8-bit integer version and H.264 (FFmpeg).

C. Fast Optimization Based on Implicit Neural Representations

There are two types of Implicit Neural Representation (INR). One uses the pixel coordinates as inputs and the network learns to generate the color values for the pixel in question. Another simply overfits an autoencoder or a decoder for a given image/video.

For the first category, starting from [304], there have been quite a few INR-based image/video compression methods [305], [306], [307], [308], [309], [310], [311], [312], [313], [314].

As one of the very first attempt at INR compression, [304] overfits the image with a small MLP. After that, the weights of MLP are quantized and stored as the bitstream. Compared with the typical hyperprior frameworks, there is still room for improvement in terms of compression performance. However, the model size is only 14 kB, which is smaller than the hyperprior by several orders of magnitude.

Reference [305] tackles some issues of the design in [304]. The first issue is the prolonged encoding time for overfitting the model. To accelerate the overfitting process, [305] uses the idea of meta learning to determine the initial weights. The second issue is the inferior coding performance. To solve this problem, the post-quantization optimization and entropy coding have been proposed for the INR-based compression framework. As a result, it outperforms [304] significantly in terms of the coding performance. The convergence speed is also much faster than [304].

As a very recent work, [306] proposes a combined manner of latent and INR-based compression. INR works for two networks: the entropy network and synthesis network. At the decoding side, the entropy network and the synthesis network are reconstructed based on the decoded weights. The means and scales are then generated to formulate the prior distribution. Through the entropy decoding, the latent is recovered. Finally, the latent will be sent to the synthesis network to generate the decoded image. In addition to image compression, [306] also extends the framework to video compression. As a result, [306] only consumes 3 kMAC/Pixel and 5 kMAC/Pixel for image and video, respectively. Furthermore, its coding performance is quite attractive. For image coding, it is comparable with VVC. For video, it is comparable with [90].

For the second category, the key concept is to use overfitting to reduce the amortization gap [315]. There have been several works [316], [317], [318], [319], [320]. Reference [318] improves the reconstruction quality by overfitting the bias in the decoder. Reference [316] fine tunes not only the encoder and latent, but also the entire model. Reference [317] makes the factorized prior and hyperprior models more suitable to the test instance. Reference [319] studies how to overfit some important parameters in order to reduce the overhead. Different from the above works for images, [320] fine tunes the model for video.

D. Other Optimization Schemes for Practicality Considerations

The above three subsections are mainly for fast and low-complexity implementations. In addition to the speed, there are several optimization techniques for practicality considerations. We mainly introduce the adversarial attack and variable bitrate in this subsection.

Adversarial attack is one way to mislead the results of neural networks. For learned codecs, since the neural network structure is usually disclosed, the attacker can easily fetch the network parameters and generate adversarial inputs. The target of the adversarial attack can be the reconstructed quality (e.g. PSNR, MS-SSIM) or bitrate. There have been several works targeting at the attack and defense of learned codecs [321], [322], [323], [324]. Reference [321] is a very early attempt at the white-box and black-box attack on learned image compression. Reference [322] proposes a training-free defense framework with a random input transform. The method does not influence the rate-distortion result for the clean image. Reference [323] gives a comprehensive study on various attack methods, attacking targets, neural network structures and bitrates. The attack transferability to VVC was also studied in [323]. Reference [324] also tries various settings for the attack. Besides, several efficient defense methods such as pre-processing and adversarial training are proposed.

Variable bitrate is another important issue. Different from the traditional codec which utilizes quantization

parameters (QPs) to control the bitrate, learned codecs usually incorporates a hyperparmeter λ to adjust the trade-off between rate and distortion. Each λ corresponds to a specific model, which increases the storage cost of network models. To solve this problem, there have been several works [325], [326], [327]. Reference [325] adopts a conditional autoencoder. λ and quantization bin size are used for the rate control. Reference [326] adopts multiple λ in the training phase, so that the resulting model is able to interpolate between the pre-trained λ to achieve variable-rate coding. Reference [327] utilizes a quality map to generate prior condition features, and then insert these features into the encoder and decoder to realize variable-rate coding.

VIII. Conclusion

Generative models have become a vital and quickly progressing field of study in the era of artificial intelligence. They have achieved great success in various tasks and have also exerted a substantial impact on visual signal coding and processing. This paper offers a concise review of generative models, along with a comprehensive survey of visual signal coding with generative models, focusing specifically on algorithms for image and video coding. Additionally, it addresses the recent advancements in international standardization efforts for visual signal coding with generative models. These efforts are extremely important for media industry. This paper also discusses the research works of applying generative models to various image and video processing tasks, such as restoration, synthesis, editing, and interpolation, along with visual signal quality assessment using generative models and quality assessment for generative models. Finally, this paper discusses the latest advancements in optimization research on visual signal coding and processing with generative models. The field of visual signal coding and processing with generative models is vast and rapidly evolving, making it difficult to undertake a comprehensive overview that includes all relevant works. Inevitably, some important research or emerging trends may have been overlooked in this paper. We hope that this survey will provide valuable insights and encourage further exploration and innovation among researchers in this field.

References

[1] I. Goodfellow, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 2672–2680.

[2] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*.

[3] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 2172–2180.

[4] Y. Pang, J. Lin, T. Qin, and Z. Chen, "Image-to-image translation: Methods and applications," *IEEE Trans. Multimedia*, vol. 24, pp. 3859–3881, 2022.

[5] C. Ledig et al., "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4681–4690.

[6] Y.-S. Chen, Y.-C. Wang, M.-H. Kao, and Y.-Y. Chuang, "Deep photo enhancer: Unpaired learning for image enhancement from photographs with GANs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 6306–6314.

[7] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 4401–4410.

[8] Y. Shen, C. Yang, X. Tang, and B. Zhou, "InterFaceGAN: Interpreting the disentangled face representation learned by GANs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 4, pp. 2004–2018, Apr. 2022.

[9] S. Tulyakov, M. Liu, X. Yang, and J. Kautz, "MoCoGAN: Decomposing motion and content for video generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1526–1535.

[10] L. Yu, W. Zhang, J. Wang, and Y. Yu, "SeqGAN: Sequence generative adversarial nets with policy gradient," in *Proc. AAAI Conf. Artif. Intell.*, 2017, vol. 31, no. 1, pp. 2852–2858.

[11] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 17022–17033.

[12] R. Li, X. Li, K.-H. Hui, and C.-W. Fu, "SP-GAN: Sphere-guided 3D shape generation and manipulation," *ACM Trans. Graph.*, vol. 40, no. 4, pp. 1–12, Aug. 2021.

[13] M. Arjovsky and L. Bottou, "Towards principled methods for training generative adversarial networks," 2017, *arXiv:1701.04862*.

[14] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 214–223.

[15] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. Int. Conf. Learn. Represent.*, 2013.

[16] Y. Burda, R. B. Grosse, and R. Salakhutdinov, "Importance weighted autoencoders," in *Proc. 4th Int. Conf. Learn. Represent. (ICLR)*, Y. Bengio and Y. LeCun, Eds., 2016.

[17] A. Razavi, A. van den Oord, B. Poole, and O. Vinyals, "Preventing posterior collapse with delta-vaes," in *Proc. 7th Int. Conf. Learn. Represent. (ICLR)*, 2019.

[18] A. Vahdat and J. Kautz, "NVAE: A deep hierarchical variational autoencoder," in *Proc. 34th Int. Conf. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 19667–19679.

[19] D. P. Kingma, T. Salimans, R. Józefowicz, X. Chen, I. Sutskever, and M. Welling, "Improving variational autoencoders with inverse autoregressive flow," in *Proc. Neural Inf. Process. Syst.*, 2016, pp. 4736–4744.

[20] R. Child, "Very deep VAEs generalize autoregressive models and can outperform them on images," in *Proc. Int. Conf. Learn. Represent.*, 2020.

[21] G. Ostrovski, W. Dabney, and R. Munos, "Autoregressive quantile networks for generative modeling," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 3936–3945.

[22] A. Van Den Oord, N. Kalchbrenner, and K. Kavukcuoglu, "Pixel recurrent neural networks," in *Proc. Int. Conf. Mach. Learn.*, Jun. 2016, pp. 1747–1756.

[23] A. Van den Oord, N. Kalchbrenner, L. Espeholt, O. Vinyals, A. Graves, and K. Kavukcuoglu, "Conditional image generation with PixelCNN decoders," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 4790–4798.

[24] S. Bond-Taylor, A. Leach, Y. Long, and C. G. Willcocks, "Deep generative modelling: A comparative review of VAEs, GANs, normalizing flows, energy-based and autoregressive models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 7327–7347, Nov. 2022.

[25] E. Hoogeboom, A. A. Gritsenko, J. Bastings, B. Poole, R. van den Berg, and T. Salimans, "Autoregressive diffusion models," in *Proc. Int. Conf. Learn. Represent.*, 2021.

[26] T. Wu et al., "AR-Diffusion: Auto-regressive diffusion model for text generation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 36, 2024, pp. 39957–39974.

[27] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 6840–6851.

[28] D. J. Rezende and S. Mohamed, "Variational inference with normalizing flows," in *Proc. Int. Conf. Mach. Learn. (ICML)*, vol. 37, 2015, pp. 1530–1538.

[29] I. Kobyzev, S. J. D. Prince, and M. A. Brubaker, "Normalizing flows: An introduction and review of current methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 11, pp. 3964–3979, Nov. 2021.

[30] W. Grathwohl, R. T. Chen, J. Bettencourt, I. Sutskever, and D. Duvenaud, "FFJORD: Free-form continuous dynamics for scalable reversible generative models," in *Proc. Int. Conf. Learn. Represent.*, 2018.

[31] L. Dinh, J. Sohl-Dickstein, and S. Bengio, "Density estimation using real nvp," in *Proc. Int. Conf. Learn. Represent.*, 2016.

[32] D. P. Kingma and P. Dhariwal, "Glow: Generative flow with invertible 1×1 convolutions," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 10236–10245.

[33] R. T. Chen, J. Behrmann, D. K. Duvenaud, and J.-H. Jacobsen, "Residual flows for invertible generative modeling," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 9913–9923.

[34] M. S. Albergo and E. Vanden-Eijnden, "Building normalizing flows with stochastic interpolants," in *Proc. 11th Int. Conf. Learn. Represent.*, 2022.

[35] Q. Zhang and Y. Chen, "Diffusion normalizing flow," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 16280–16291.

[36] M. Zand, A. Etemad, and M. Greenspan, "Diffusion models with deterministic normalizing flow priors," 2023, *arXiv:2309.01274*.

[37] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *Proc. Int. Conf. Learn. Represent.*, 2020.

[38] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *Proc. Int. Conf. Learn. Represent.*, 2020.

[39] C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, and J. Zhu, "DPM-Solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 5775–5787.

[40] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 10684–10695.

[41] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, and L. van Gool, "RePaint: Inpainting using denoising diffusion probabilistic models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11461–11471.

[42] G. Parmar, K. K. Singh, R. Zhang, Y. Li, J. Lu, and J.-Y. Zhu, "Zero-shot image-to-image translation," in *Proc. ACM SIGGRAPH Conf.*, 2023, pp. 1–11.

[43] S. Lu, Y. Liu, and A. W.-K. Kong, "TF-ICON: Diffusion-based training-free cross-domain image composition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 2294–2305.

[44] R. Gal et al., "An image is worth one word: Personalizing text-to-image generation using textual inversion," in *Proc. 11th Int. Conf. Learn. Represent.*, 2022.

[45] L. Zhang and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2023, pp. 3836–3847.

[46] A. Blattmann et al., "Stable video diffusion: Scaling latent video diffusion models to large datasets," 2023, *arXiv:2311.15127*.

[47] B. Poole, A. Jain, J. T. Barron, and B. Mildenhall, "DreamFusion: Text-to-3D using 2D diffusion," in *Proc. 11th Int. Conf. Learn. Represent.*, 2022.

[48] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2256–2265.

[49] J. Ballé, V. Laparra, and E. P. Simoncelli, "End-to-end optimized image compression," in *Proc. Int. Conf. Learn. Representat.*, 2016.

[50] L. Theis, T. Salimans, M. D. Hoffman, and F. Mentzer, "Lossy compression with Gaussian diffusion," 2022, *arXiv:2206.08889*.

[51] D. Minnen, J. Ballé, and G. D. Toderici, "Joint autoregressive and hierarchical priors for learned image compression," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 10794–10803.

[52] Y. Xie, K. L. Cheng, and Q. Chen, "Enhanced invertible encoding for learned image compression," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 162–170.

[53] Z. Guo, Z. Zhang, R. Feng, and Z. Chen, "Soft then hard: Rethinking the quantization in neural image compression," in *Proc. 38th Int. Conf. Mach. Learn.*, Jul. 2021, pp. 3920–3929.

[54] J. Ballé, D. Minnen, S. Singh, S. J. Hwang, and N. Johnston, "Variational image compression with a scale hyperprior," in *Proc. Int. Conf. Learn. Represent.*, 2018.

[55] Z. Guo, Z. Zhang, R. Feng, and Z. Chen, "Causal contextual prediction for learned image compression," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 4, pp. 2329–2341, Apr. 2022.

[56] D. He, Z. Yang, W. Peng, R. Ma, H. Qin, and Y. Wang, "ELIC: Efficient learned image compression with unevenly grouped space-channel contextual adaptive coding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5718–5727.

[57] F. Mentzer, G. D. Toderici, M. Tschannen, and E. Agustsson, "High-fidelity generative image compression," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 11913–11924.

[58] E. Agustsson et al., "Soft-to-hard vector quantization for end-to-end learning compressible representations," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1141–1151.

[59] R. Feng, Z. Guo, W. Li, and Z. Chen, "NVTC: Nonlinear vector transform coding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 6101–6110.

[60] Y. Yang, R. Bamler, and S. Mandt, "Improving inference for neural image compression," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 573–584.

[61] D. Kingma, T. Salimans, B. Poole, and J. Ho, "Variational diffusion models," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 21696–21707.

[62] G. Flamich, M. Havasi, and J. M. Hernández-Lobato, "Compressing images by encoding their latent representations with relative entropy coding," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 16131–16141.

[63] G. Flamich, S. Markou, and J. M. Hernández-Lobato, "Fast relative entropy coding with A* coding," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 6548–6577.

[64] D. Minnen and S. Singh, "Channel-wise autoregressive entropy models for learned image compression," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2020, pp. 3339–3343.

[65] Y. Li, T. Xu, Y. Wang, J. Liu, and Y.-Q. Zhang, "Idempotent learned image compression with right-inverse," in *Proc. 37th Conf. Neural Inf. Process. Syst.*, 2023, pp. 12878–12896.

[66] L. Helminger, A. Djelouah, M. Gross, and C. Schroers, "Lossy image compression with normalizing flows," in *Proc. Neural Compression Inf. Theory Appl. Workshop @ ICLR*, 2021.

[67] Y.-H. Ho, C.-C. Chan, W.-H. Peng, H.-M. Hang, and M. Domanski, "ANFIC: Image compression using augmented normalizing flows," *IEEE Open J. Circuits Syst.*, vol. 2, pp. 613–626, 2021.

[68] Y.-H. Ho, C.-P. Chang, P.-Y. Chen, A. Gnutti, and W.-H. Peng, "CANF-VC: Conditional augmented normalizing flows for video compression," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Oct. 2022, pp. 207–223.

[69] M.-J. Chen, Y.-H. Chen, and W.-H. Peng, "B-CANF: Adaptive B-frame coding with conditional augmented normalizing flows," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 4, pp. 2908–2921, Apr. 2024.

[70] D. Alexandre, H.-M. Hang, and W.-H. Peng, "Hierarchical B-frame video coding using two-layer CANF without motion coding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 10249–10258.

[71] H. Ma, D. Liu, R. Xiong, and F. Wu, "IWave: CNN-based wavelet-like transform for image compression," *IEEE Trans. Multimedia*, vol. 22, no. 7, pp. 1667–1679, Jul. 2020.

[72] E. Hoogeboom, J. Peters, R. Van Den Berg, and M. Welling, "Integer discrete flows and lossless compression," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 12134–12144.

[73] R. van den Berg, A. A. Gritsenko, M. Dehghani, C. Kaae Sønderby, and T. Salimans, "IDF++: Analyzing and improving integer discrete flows for lossless compression," 2020, *arXiv:2006.12459*.

[74] Y. Blau and T. Michaeli, "The perception-distortion tradeoff," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6228–6237.

[75] E. Agustsson, M. Tschannen, F. Mentzer, R. Timofte, and L. Van Gool, "Generative adversarial networks for extreme learned image compression," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 221–231.

[76] E. Agustsson, D. Minnen, G. Toderici, and F. Mentzer, "Multi-realism image compression with a conditional generator," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 22324–22333.

[77] R. Yang and S. Mandt, "Lossy image compression with conditional diffusion models," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 36, 2024, pp. 64971–64995.

[78] N. F. Ghouse, J. Petersen, A. Wiggers, T. Xu, and G. Sautière, "A residual diffusion model for high perceptual quality codec augmentation," 2023, *arXiv:2301.05489*.

[79] E. Hoogeboom, E. Agustsson, F. Mentzer, L. Versari, G. Toderici, and L. Theis, "High-fidelity image compression with score-based generative models," 2023, *arXiv:2305.18231*.

[80] Y. Ye and M. Wien, *Call for Learning-based Video Codecs for Study of Quality Assessment*, ISO/IEC Standard JTC 1/SC 29/AG 5 N104, 2023.

[81] C. E. Shannon et al., "Coding theorems for a discrete source with a fidelity criterion," *IRE Nat. Conv. Rec.*, vol. 4, no. 4, pp. 142–163, 1959.

[82] Y. Blau and T. Michaeli, "Rethinking lossy compression: The rate-distortion-perception tradeoff," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 675–685.

[83] Z. Yan, F. Wen, R. Ying, C. Ma, and P. Liu, "On perceptual lossy compression: The cost of perceptual reconstruction and an optimal training framework," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 11682–11692.

[84] G. Zhang, J. Qian, J. Chen, and A. Khisti, "Universal rate-distortion-perception representations for lossy compression," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 11517–11529.

[85] D. Freirich, T. Michaeli, and R. Meir, "A theory of the distortion-perception tradeoff in Wasserstein space," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 25661–25672.

[86] J. Chen, L. Yu, J. Wang, W. Shi, Y. Ge, and W. Tong, "On the rate-distortion-perception function," *IEEE J. Sel. Areas Inf. Theory*, vol. 3, no. 4, pp. 664–673, Dec. 2022.

[87] T. Xu et al., "Idempotence and perceptual image compression," in *Proc. Int. Conf. Learn. Represent.*, 2024.

[88] A. Habibian, T. V. Rozendaal, J. Tomczak, and T. Cohen, "Video compression with rate-distortion autoencoders," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7033–7042.

[89] J. Liu et al., "Conditional entropy coding for efficient video compression," in *Proc. 16th Eur. Conf. Comput. Vis. (ECCV)*, Glasgow, U.K. Springer, Aug. 2020, pp. 453–468.

[90] F. Mentzer et al., "VCT: A video compression transformer," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 13091–13103.

[91] Z. Chen, T. He, X. Jin, and F. Wu, "Learning for video compression," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 2, pp. 566–576, Feb. 2020.

[92] O. Rippel, S. Nair, C. Lew, S. Branson, A. G. Anderson, and L. Bourdev, "Learned video compression," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3454–3463.

[93] G. Lu, W. Ouyang, D. Xu, X. Zhang, C. Cai, and Z. Gao, "DVC: An end-to-end deep video compression framework," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11006–11015.

[94] Z. Hu, Z. Chen, D. Xu, G. Lu, W. Ouyang, and S. Gu, "Improving deep video compression by resolution-adaptive flow coding," in *Proc. Eur. Conf. Comput. Vis.*, Glasgow, U.K. Springer, Aug. 2020, pp. 193–209.

[95] E. Agustsson, D. Minnen, N. Johnston, J. Balle, S. J. Hwang, and G. Toderici, "Scale-space flow for end-to-end optimized video compression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 8503–8512.

[96] J. Lin, D. Liu, H. Li, and F. Wu, "M-LVC: Multiple frames prediction for learned video compression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3546–3554.

[97] Z. Hu, G. Lu, and D. Xu, "FVC: A new framework towards deep video compression in feature space," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 1502–1511.

[98] G. Lu, X. Zhang, W. Ouyang, L. Chen, Z. Gao, and D. Xu, "An end-to-end learning framework for video compression," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3292–3308, Oct. 2021.

[99] R. Yang, F. Mentzer, L. Van Gool, and R. Timofte, "Learning for video compression with recurrent auto-encoder and recurrent probability model," *IEEE J. Sel. Topics Signal Process.*, vol. 15, no. 2, pp. 388–401, Feb. 2021.

[100] Z. Hu, G. Lu, J. Guo, S. Liu, W. Jiang, and D. Xu, "Coarse-to-fine deep video coding with hyperprior-guided mode prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5921–5930.

[101] Z. Guo, R. Feng, Z. Zhang, X. Jin, and Z. Chen, "Learning cross-scale weighted prediction for efficient neural video compression," *IEEE Trans. Image Process.*, vol. 32, pp. 3567–3579, 2023.

[102] C.-Y. Wu, N. Singhal, and P. Krahenbuhl, "Video compression through image interpolation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 416–431.

[103] A. Djelouah, J. Campos, S. Schaub-Meyer, and C. Schroers, "Neural inter-frame compression for video coding," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6421–6429.

[104] Z. Cheng, H. Sun, M. Takeuchi, and J. Katto, "Learning image and video compression through spatial–temporal energy compaction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10063–10072.

[105] R. Yang, F. Mentzer, L. Van Gool, and R. Timofte, "Learning for video compression with hierarchical quality and recurrent enhancement," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6628–6637.

[106] R. Pourreza and T. Cohen, "Extending neural P-frame codecs for B-frame coding," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 6680–6689.

[107] M. A. Yilmaz and A. M. Tekalp, "End-to-end rate-distortion optimized learned hierarchical bi-directional video compression," *IEEE Trans. Image Process.*, vol. 31, pp. 974–983, 2022.

[108] R. Yang, R. Timofte, and L. Van Gool, "Advancing learned video compression with in-loop frame prediction," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 5, pp. 2410–2423, May 2023.

[109] T. Ladune, P. Philippe, W. Hamidouche, L. Zhang, and O. Déforges, "ModeNet: Mode selection network for learned video coding," in *Proc. IEEE 30th Int. Workshop Mach. Learn. Signal Process. (MLSP)*, Sep. 2020, pp. 1–6.

[110] T. Ladune, P. Philippe, W. Hamidouche, L. Zhang, and O. Déforges, "Optical flow and mode selection for learning-based video coding," in *Proc. IEEE 22nd Int. Workshop Multimedia Signal Process. (MMSP)*, Sep. 2020, pp. 1–6.

[111] O. Rippel, A. G. Anderson, K. Tatwawadi, S. Nair, C. Lytle, and L. Bourdev, "ELF-VC: Efficient learned flexible-rate video coding," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 14479–14488.

[112] J. Li, B. Li, and Y. Lu, "Deep contextual video compression," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 34, Dec. 2021, pp. 18114–18125.

[113] F. Brand, J. Seiler, and A. Kaup, "On benefits and challenges of conditional interframe video coding in light of information theory," in *Proc. Picture Coding Symp. (PCS)*, Dec. 2022, pp. 289–293.

[114] X. Sheng, J. Li, B. Li, L. Li, D. Liu, and Y. Lu, "Temporal context mining for learned video compression," *IEEE Trans. Multimedia*, vol. 25, pp. 7311–7322, 2023.

[115] J. Li, B. Li, and Y. Lu, "Hybrid spatial–temporal entropy modelling for neural video compression," in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 1503–1511.

[116] J. Xiang, K. Tian, and J. Zhang, "MIMT: Masked image modeling transformer for video compression," in *Proc. 11th Int. Conf. Learn. Represent.*, 2022.

[117] L. Qi, J. Li, B. Li, H. Li, and Y. Lu, "Motion information propagation for neural video compression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 6111–6120.

[118] J. Li, B. Li, and Y. Lu, "Neural video compression with diverse contexts," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 22616–22626.

[119] F. Brand, J. Seiler, and A. Kaup, "Conditional residual coding: A remedy for bottleneck problems in conditional inter frame coding," *IEEE Trans. Circuits Syst. Video Technol.*, to be published.

[120] R. Yang, R. Timofte, and L. Van Gool, "Perceptual learned video compression with recurrent conditional GAN," in *Proc. 31st Int. Joint Conf. Artif. Intell.*, Vienna, Austria, L. D. Raedt, Ed., Jul. 2022, pp. 1537–1544.

[121] F. Mentzer, E. Agustsson, J. Ballé, D. Minnen, N. Johnston, and G. Toderici, "Neural video compression using GANs for detail synthesis and propagation," in *Proc. Eur. Conf. Comput. Vis.* Tel Aviv, Israel: Springer, Nov. 2022, pp. 562–578.

[122] B. Phan, S. Salehkalaibar, J. Chen, W. Yu, and A. J. Khisti, "On the choice of perception loss function for learned video compression," in *Proc. ICML Workshop Neural Compression Inf. Theory Appl.*, 2023, pp. 48226–48274.

[123] R. Yang, P. Srivastava, and S. Mandt, "Diffusion probabilistic modeling for video generation," *Entropy*, vol. 25, no. 10, p. 1469, Oct. 2023.

[124] E. Alshina, *JPEG AI Overview Slide*, document JPEG AI ISO/IEC JTC 1/SC29/WG1 M99064, 2023.

[125] Z. Zhang, S. Esenlik, Y. Wu, M. Wang, K. Zhang, and L. Zhang, "End-to-end learning-based image compression with a decoupled framework," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 5, pp. 3067–3081, May 2024.

[126] S. Esenlik, Y. Li, Y. Wu, K. Zhang, L. Zhang, and Z. Zhang, *Bytedance's Response to the JPEG AI Call for Proposals*, document JPEG AI ISO/IEC JTC 1/SC29/WG1 M96053, 2023.

[127] G. K. Wallace, "The JPEG still picture compression standard," *IEEE Trans. Consum. Electron.*, vol. 38, no. 1, pp. 18–34, Feb. 1992.

[128] A. Skodras, C. Christopoulos, and T. Ebrahimi, "The JPEG 2000 still image compression standard," *IEEE Signal Process. Mag.*, vol. 18, no. 5, pp. 36–58, 2001.

[129] T. Wiegand, G. J. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 7, pp. 560–576, Jul. 2003.

[130] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1649–1668, Dec. 2012.

[131] J. Han et al., "A technical overview of AV1," *Proc. IEEE*, vol. 109, no. 9, pp. 1435–1462, Sep. 2021.

[132] B. Bross et al., "Overview of the versatile video coding (VVC) standard and its applications," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 10, pp. 3736–3764, Oct. 2021.

[133] I.-K. Kim, J. Min, T. Lee, W.-J. Han, and J. Park, "Block partitioning structure in the HEVC standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1697–1706, Dec. 2012.

[134] J. Pfaff et al., "Intra prediction and mode coding in VVC," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 10, pp. 3834–3847, Oct. 2021.

[135] X. Zhao et al., "Transform coding in the VVC standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 10, pp. 3878–3890, Oct. 2021.

[136] H. Schwarz et al., "Quantization and entropy coding in the versatile video coding (VVC) standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 10, pp. 3891–3906, Oct. 2021.

[137] Z. Cheng, H. Sun, M. Takeuchi, and J. Katto, "Learned image compression with discretized Gaussian mixture likelihoods and attention modules," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7939–7948.

[138] J. Ascenso, E. Alshina, and T. Ebrahimi, "The JPEG AI standard: Providing efficient human and machine visual data consumption," *IEEE MultimediaMag.*, vol. 30, no. 1, pp. 100–111, Jan. 2023.

[139] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Proc. 37th Asilomar Conf. Signals, Syst. Comput.*, vol. 2, Nov. 2003, pp. 1398–1402.

[140] A. Lucas, S. López-Tapia, R. Molina, and A. K. Katsaggelos, "Generative adversarial networks and perceptual losses for video super-resolution," *IEEE Trans. Image Process.*, vol. 28, no. 7, pp. 3312–3327, Jul. 2019.

[141] *JPEG AI Common Training and Test Conditions*, document JPEG AI (ISO/IEC 6048) ISO/IEC JTC 1/SC29/WG1 N100106, WG1, 2022.

[142] R. Rassool, "VMAF reproducibility: Validating a perceptual practical video quality metric," in *Proc. IEEE Int. Symp. Broadband Multimedia Syst. Broadcast. (BMSB)*, Jun. 2017, pp. 1–2.

[143] N. Ponomarenko, F. Silvestri, K. Egiazarian, M. Carli, J. Astola, and V. Lukin, "On between-coefficient contrast masking of DCT basis functions," in *Proc. 3rd Int. Workshop Video Process. Qual. Metrics Consum. Electron. (VPQM)*, Scottsdale, AZ, USA, Jan. 2007, p. 4.

[144] (2021). *VVCSoftware_VTM Repository*. [Online]. Available: https://vcgit.hhi.fraunhofer.de/jvet/VVCSoftware_VTM/-/tree/VTM-11.1?ref_type=tags

[145] A. Karabutov, Y. Wu, E. Alshina, and J. Ascenso, *JPEG AI Sw V4.x Status*, document JPEG AI ISO/IEC JTC 1/SC29/WG1 M101081, 2023.

[146] Y. Zhang, K. Li, K. Li, B. Zhong, and Y. Fu, "Residual non-local attention networks for image restoration," 2019, *arXiv:1903.10082*.

[147] X. Wang et al., *Report of 3.11-Acceleration of Context Modules*, document JPEG AI ISO/IEC JTC 1/SC29/WG1 M100658, 2023.

[148] G. J. Sullivan and J.-R. Ohm, *Meeting Report of the 19th JVET Meeting (Teleconference, 22 June–1 July 2020)*, document JVET-S2000, The Joint Video Experts Team ITU-T SG 16 WP 3 ISO/IEC JTC 1/SC 29, Jun. 2020.

[149] T. Dumas, F. Galpin, P. Bordes, and F. L. Léannec, *AHG11: Neural Network-Based Intra Prediction With Transform Selection in VVC*, document JVET-T0073, The Joint Video Experts Team ITU-T SG 16 WP 3 ISO/IEC JTC 1/SC 29, Oct. 2020.

[150] T. Dumas, F. Galpin, and P. Bordes, *EE1-3.2: Neural Network-Based Intra Prediction With Learned Mapping to VVC Intra Prediction Modes*, document JVET-AC0116, The Joint Video Experts Team ITU-T SG 16 WP 3 ISO/IEC JTC 1/SC 29, Jan. 2023.

[151] T. Dumas, F. Galpin, and P. Bordes, *Non-EE1: Neural Network-Based Intra Prediction With Learned Mapping to VVC Intra Prediction Modes*, document JVET-AB0149, The Joint Video Experts Team ITU-T SG 16 WP 3 ISO/IEC JTC 1/SC 29, Oct. 2022.

[152] T. Dumas, F. Galpin, and P. Bordes, *AHG11: Neural Network-Based Intra Prediction With Reduced Complexity*, document JVET-AD0212, The Joint Video Experts Team ITU-T SG 16 WP 3 ISO/IEC JTC 1/SC 29, Apr. 2023.

[153] Y. Chu, Z. Wang, W. Zhang, and S. Li, *AHG11: Neural Network-Based Reference Cu Quality Enhancement for Motion Compensation Prediction*, document JVET-AC0090, The Joint Video Experts Team ITU-T SG 16 WP 3 ISO/IEC JTC 1/SC 29, Jan. 2023.

[154] J. Jia et al., *AHG11: Deep Reference Frame Generation for Inter Prediction Enhancement*, document JVET-AC0090, The Joint Video Experts Team ITU-T SG 16 WP 3 ISO/IEC JTC 1/SC 29, Jan. 2023.

[155] H. Wang, M. Karczewicz, J. Chen, and A. Kotra, *AHG11: Neural Network-Based In-Loop Filter*, document JVET-T0079, The Joint Video Experts Team ITU-T SG 16 WP 3 ISO/IEC JTC 1/SC 29, Oct. 2020.

[156] Y. Li, L. Zhang, and K. Zhang, *AHG11: Convolutional Neural Networks-Based In-Loop Filter*, document JVET-T0088, The Joint Video Experts Team ITU-T SG 16 WP 3 ISO/IEC JTC 1/SC 29, Oct. 2020.

[157] D. Liu, J. Ström, M. Damghanian, P. Wennersten, and K. Andersson, *EE1-1.5: Combined Intra and Inter Models for Luma and Chroma*, document JVET-AC0089, The Joint Video Experts Team ITU-T SG 16 WP 3 ISO/IEC JTC 1/SC 29, Jan. 2023.

[158] J. N. Shingala et al., *EE1-1.1: Complexity Reduction on Neural-Network Loop Filter*, document JVET-AD0156, The Joint Video Experts Team ITU-T SG 16 WP 3 ISO/IEC JTC 1/SC 29, Apr. 2023.

[159] D. Rusanovskyy et al., *AHG11/EE1: Status of the Joint EE1-0 (LOP.2) Training*, document JVET-AF0043, The Joint Video Experts Team ITU-T SG 16 WP 3 ISO/IEC JTC 1/SC 29, Oct. 2023.

[160] Y. Li et al., *AHG11: Content-Adaptive Neural Network Post-Processing Filter*, document JVET-V0075, The Joint Video Experts Team ITU-T SG 16 WP 3 ISO/IEC JTC 1/SC 29, Apr. 2021.

[161] J. Y. Lee, Y. Choi, W. Lim, and G. Bang, *AHG11: Deep Neural Network for Super-Resolution*, document JVET-T0096, The Joint Video Experts Team ITU-T SG 16 WP 3 ISO/IEC JTC 1/SC 29, Oct. 2020.

[162] R. Chang, L. Wang, X. Xu, and S. Liu, *EE1-2.2: GOP Level Adaptive Resampling With CNN-Based Super Resolution*, document JVET-AC0196, The Joint Video Experts Team ITU-T SG 16 WP 3 ISO/IEC JTC 1/SC 29, Jan. 2023.

[163] Y. He, B. Wang, E. Alshina, and J. Sauer, *AHG11: A Hybrid Codec Using E2E Image Coding Combined With VVC Video Coding*, document JVET-AA0063, The Joint Video Experts Team ITU-T SG 16 WP 3 ISO/IEC JTC 1/SC 29, Jul. 2022.

[164] S. Jiang, Z. Wang, W. Zhang, and S. Li, *AHG11: Fourier Series and Laplacian Noise-Based Quantization Error Compensation for End-to-End Learning-Based Image Compression*, document JVET-AC0091, The Joint Video Experts Team ITU-T SG 16 WP 3 ISO/IEC JTC 1/SC 29, Jan. 2023.

[165] T. Dumas, F. Galpin, and P. Bordes, "Iterative training of neural networks for intra prediction," *IEEE Trans. Image Process.*, vol. 30, pp. 697–711, 2021.

[166] Y. Li et al., *EE1-1.6: Deep In-Loop Filter With Fixed Point Implementation*, document JVET-AA0111, The Joint Video Experts Team ITU-T SG 16 WP 3 ISO/IEC JTC 1/SC 29, Jul. 2022.

[167] J. Li, Y. Li, K. Zhang, and L. Zhang, *EE1-1.6: RDO Considering Deep In-Loop Filtering*, document JVET-AB0068, The Joint Video Experts Team ITU-T SG 16 WP 3 ISO/IEC JTC 1/SC 29, Oct. 2022.

[168] Y. Li, K. Zhang, and L. Zhang, *EE1-1.7: Deep In-Loop Filter With Additional Input Information*, document JVET-AC0177, The Joint Video Experts Team ITU-T SG 16 WP 3 ISO/IEC JTC 1/SC 29, Jan. 2023.

[169] Y. Li, L. Zhang, and K. Zhang, "IDAM: Iteratively trained deep in-loop filter with adaptive model selection," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 19, no. 1s, pp. 1–22, Feb. 2023.

[170] F. Galpin, Y. Li, Y. Li, J. N. Shingala, L. Wang, and Z. Xie, *Report: NNVC Software Development AhG14*, document JVET-AG0014, Joint Video Experts Team (JVET) ITU-T SG 16 WP 3 ISO/IEC JTC 1/SC 29, Jan. 2024.

[171] W. F. Schreiber, C. F. Knapp, and N. D. Kay, "Synthetic highs—An experimental TV bandwidth reduction system," *J. SMPTE*, vol. 68, no. 8, pp. 525–537, Aug. 1959.

[172] D. E. Pearson and J. A. Robinson, "Visual communication at very low data rates," *Proc. IEEE*, vol. 73, no. 4, pp. 795–812, Apr. 1985.

[173] R. Lopez and T. Huang, "Head pose computation for very low bit-rate video coding," in *Proc. Int. Conf. Comput. Anal. Images Patterns*, 1995, pp. 440–447.

[174] Y. Nakaya, Y. C. Chuah, and H. Harashima, "Model-based/waveform hybrid coding for videotelephone images," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Jan. 1991, pp. 2741–2744.

[175] M. Oquab et al., "Low bandwidth video-chat compression using deep generative models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 2388–2397.

[176] G. Konuko, G. Valenzise, and S. Lathuilière, "Ultra-low bitrate video conferencing using deep image animation," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2022, pp. 3515–3520.

[177] A. Tang et al., "Generative compression for face video: A hybrid scheme," in *Proc. IEEE Int. Conf. Multimedia Expo. (ICME)*, Jul. 2022, pp. 1–6.

[178] B. Chen, Z. Wang, B. Li, R. Lin, S. Wang, and Y. Ye, "Beyond keypoint coding: Temporal evolution inference with compact feature representation for talking face video compression," in *Proc. Data Compress. Conf. (DCC)*, Mar. 2022, pp. 13–22.

[179] Z. Wang, B. Chen, Y. Ye, and S. Wang, "Dynamic multi-reference generative prediction for face video compression," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2022, pp. 896–900.

[180] B. Chen, Z. Wang, B. Li, S. Wang, and Y. Ye, "Compact temporal trajectory representation for talking face video compression," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 11, pp. 7009–7023, Nov. 2023.

[181] B. Chen, Z. Wang, B. Li, S. Wang, S. Wang, and Y. Ye, "Interactive face video coding: A generative compression framework," 2023, *arXiv:2302.09919*.

[182] S. Yin et al., "Enabling translatability of generative face video coding: A unified face feature transcoding framework," in *Proc. Data Compress. Conf.*, 2024, pp. 113–122.

[183] P. Dhariwal and A. Nichol, "Diffusion models beat GANs on image synthesis," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 8780–8794.

[184] B. Chen et al., *AHG9: Common SEI Message of Generative Face Video*, document JVET-AD0051, The Joint Video Experts Team of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29, Apr. 2023.

[185] H.-B. Teo et al., *AHG9: A Study on Generative Face Video SEI Message*, document JVET-AE0088, The Joint Video Experts Team of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29, Jul. 2023.

[186] M. M. Hannuksela et al., *AHG9: On the Generative Face Video SEI Message*, document JVET-AG0087, The Joint Video Experts Team of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29, Jan. 2024.

[187] B. Chen et al., *AHG9: Common Text for Proposed Generative Face Video SEI Message*, document JVET-AE0280, The Joint Video Experts Team of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29, Jul. 2023.

[188] M. M. Hannuksela et al., *AHG9: Usage of the Neural-Network Post-Filter Characteristics Sei Message to Define the Generator NN of the Generative Face Video SEI Message*, document JVET-AG0088, The Joint Video Experts Team of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29, Jan. 2024.

[189] Y. Ye et al., *On VVC-assisted Ultra-low Rate Generative Face Video Coding*, document M64987, MPEG ISO/IEC JTC 1/SC 29/WG 2, Oct. 2023.

[190] B. Chen et al., *AHG16: Proposed Common Software Tools and Testing Conditions for Generative Face Video Compression*, document JVET-AG0042, The Joint Video Experts Team of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29, Jan. 2024.

[191] S. McCarthy and B. Chen, *Test Conditions and Evaluation Procedures for Generative Face Video Coding*, document JVET-AG2035, The Joint Video Experts Team ITU-T SG 16 WP 3 ISO/IEC JTC 1/SC 29, Jan. 2024.

[192] S. Yin et al., *AHG16: Interoperability Study on Parameter Translator of Generative Face Video Coding*, document JVET-AG0048, The Joint Video Experts Team of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29, Jan. 2024.

[193] R. Zou et al., *AHG16: Depthwise Separable Convolution for Generative Face Video Compression*, document JVET-AG0139, The Joint Video Experts Team of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29, Jan. 2024.

[194] N. Anantrasirichai and D. Bull, "Artificial intelligence in the creative industries: A review," *Artif. Intell. Rev.*, vol. 55, no. 1, pp. 589–656, Jan. 2022.

[195] C. Rota, M. Buzzelli, S. Bianco, and R. Schettini, "Video restoration based on deep learning: A comprehensive survey," *Artif. Intell. Rev.*, vol. 56, no. 6, pp. 5317–5364, Jun. 2023.

[196] L. Zhai, Y. Wang, S. Cui, and Y. Zhou, "A comprehensive review of deep learning-based real-world image restoration," *IEEE Access*, vol. 11, pp. 21049–21067, 2023.

[197] K. Chauhan et al., "Deep learning-based single-image super-resolution: A comprehensive review," *IEEE Access*, vol. 11, pp. 21811–21830, 2023.

[198] X. Li et al., "Diffusion models for image restoration and enhancement—A comprehensive survey," 2023, *arXiv:2308.09388*.

[199] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2014, pp. 1278–1286.

[200] D. Im, S. Ahn, R. Memisevic, and Y. Bengio, "Denoising criterion for variational auto-encoding framework," in *Proc. AAAI Conf. Artif. Intell.*, 2017, vol. 31, no. 1, pp. 2059–2065.

[201] A. Creswell and A. A. Bharath, "Denoising adversarial autoencoders," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 4, pp. 968–984, Apr. 2019.

[202] M. Prakash, A. Krull, and F. Jug, "Fully unsupervised diversity denoising with convolutional variational autoencoders," 2020, *arXiv:2006.06072*.

[203] T. M. Nimisha, A. K. Singh, and A. N. Rajagopalan, "Blur-invariant deep learning for blind-deblurring," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4762–4770.

[204] M. Asim, F. Shamshad, and A. Ahmed, "Blind image deconvolution using deep generative priors," *IEEE Trans. Comput. Imag.*, vol. 6, pp. 1493–1506, 2020.

[205] Y. Du, J. Xu, Q. Qiu, X. Zhen, and L. Zhang, "Variational image deraining," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, Mar. 2020, pp. 2406–2415.

[206] G. Kim, S. W. Park, and J. Kwon, "Pixel-wise Wasserstein autoencoder for highly generative dehazing," *IEEE Trans. Image Process.*, vol. 30, pp. 5452–5462, 2021.

[207] Z.-S. Liu, W.-C. Siu, and Y.-L. Chan, "Photo-realistic image super-resolution via variational autoencoders," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 4, pp. 1351–1365, Apr. 2021.

[208] D. Chira, I. Haralampiev, O. Winther, A. Dittadi, and V. Liévin, "Image super-resolution with deep variational autoencoders," in *Proc. Eur. Conf. Comput. Vis.* Tel Aviv, Israel: Springer, 2022, pp. 395–411.

[209] J. Chen, J. Chen, H. Chao, and M. Yang, "Image blind denoising with generative adversarial network based noise modeling," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3155–3164.

[210] N. Anantrasirichai and D. Bull, "Contextual colorization and denoising for low-light ultra high resolution sequences," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2021, pp. 1614–1618.

[211] Y. Poirier-Ginter and J.-F. Lalonde, "Robust unsupervised StyleGAN image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 22292–22301.

[212] S. Cha, T. Park, B. Kim, J. Baek, and T. Moon, "GAN2GAN: Generative noise learning for blind denoising with single noisy images," in *Proc. Int. Conf. Learn. Represent.*, 2021.

[213] R. Li, J. Pan, Z. Li, and J. Tang, "Single image dehazing via conditional generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8202–8211.

[214] H. Zhang, V. Sindagi, and V. M. Patel, "Image de-raining using a conditional generative adversarial network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 11, pp. 3943–3956, Nov. 2020.

[215] M. Fu, H. Liu, Y. Yu, J. Chen, and K. Wang, "DW-GAN: A discrete wavelet transform GAN for nonhomogeneous dehazing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 203–212.

[216] K. Lin, T. H. Li, S. Liu, and G. Li, "Real photographs denoising with noise domain adaptation and attentive generative adversarial network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 1717–1721.

[217] Y. Xie, E. Franz, M. Chu, and N. Thuerey, "TempoGAN: A temporally coherent, volumetric GAN for super-resolution fluid flow," *ACM Trans. Graph.*, vol. 37, no. 4, pp. 1–15, Aug. 2018.

[218] A. Kudo, Y. Kitamura, Y. Li, S. Iizuka, and E. Simo-Serra, "Virtual thin slice: 3D conditional GAN-based super-resolution for CT slice interval," in *Proc. 2nd Int. Workshop Mach. Learn. Med. Image Reconstruction (MLMIR)*, Shenzhen, China. Springer, Oct. 17, 2019, pp. 91–100.

[219] S. Maeda, "Unpaired image super-resolution using pseudo-supervision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 288–297.

[220] J. Wang, Y. Chen, Y. Wu, J. Shi, and J. Gee, "Enhanced generative adversarial network for 3D brain MRI super-resolution," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 3627–3636.

[221] X. Wang et al., "ESRGAN: Enhanced super-resolution generative adversarial networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV) Workshops*, vol. 11133. Munich, Germany, Sep. 2018, pp. 63–79.

[222] D. Ma, M. Afonso, F. Zhang, and D. R. Bull, "Perceptually-inspired super-resolution of compressed videos," *Proc. SPIE*, vol. 11137, pp. 310–318, Sep. 2019.

[223] T. Shang, Q. Dai, S. Zhu, T. Yang, and Y. Guo, "Perceptual extreme super resolution network with receptive field block," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 1778–1787.

[224] C. You et al., "CT super-resolution GAN constrained by the identical, residual, and cycle learning ensemble (GAN-CIRCLE)," *IEEE Trans. Med. Imag.*, vol. 39, no. 1, pp. 188–203, Jan. 2020.

[225] J. Park, S. Son, and K. M. Lee, "Content-aware local GAN for photo-realistic super-resolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 10585–10594.

[226] J. He, W. Shi, K. Chen, L. Fu, and C. Dong, "GCFSR: A generative and controllable face super resolution method without facial and GAN priors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 1889–1898.

[227] F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah, "Diffusion models in vision: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 9, pp. 10850–10869, Sep. 2023.

[228] O. Özdenizci and R. Legenstein, "Restoring vision in adverse weather conditions with patch-based denoising diffusion models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 8, pp. 10346–10357, Aug. 2023.

[229] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi, "Image super-resolution via iterative refinement," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 4, pp. 4713–4726, Apr. 2023.

[230] H. Li et al., "SRDiff: Single image super-resolution with diffusion probabilistic models," *Neurocomputing*, vol. 479, pp. 47–59, Mar. 2022.

[231] B. Kawar, M. Elad, S. Ermon, and J. Song, "Denoising diffusion restoration models," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 23593–23606.

[232] S. Gao et al., "Implicit diffusion models for continuous super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 10021–10030.

[233] B. Xia et al., "DiffIR: Efficient diffusion model for image restoration," 2023, *arXiv:2303.09472*.

[234] F. Zhan et al., "Multimodal image synthesis and editing: The generative AI era," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 12, pp. 15098–15119, Dec. 2023.

[235] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional GANs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8798–8807.

[236] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 2337–2346.

[237] W. Wang et al., "Semantic image synthesis via diffusion models," 2022, *arXiv:2207.00050*.

[238] C. H. Wu and F. De La Torre, "A latent space of stochastic diffusion models for zero-shot image editing and guidance," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 7378–7387.

[239] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. Singh, and M.-H. Yang, "Diverse image-to-image translation via disentangled representations," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 35–51.

[240] V. Sushko, E. Schönfeld, D. Zhang, J. Gall, B. Schiele, and A. Khoreva, "You only need adversarial supervision for semantic image synthesis," 2020, *arXiv:2012.04781*.

[241] M. Chen et al., "Generative pretraining from pixels," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1691–1703.

[242] P. Esser, R. Rombach, and B. Ommer, "Taming transformers for high-resolution image synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12873–12883.

[243] C. Wu et al., "Nüwa: Visual synthesis pre-training for neural visual world creation," in *Proc. Eur. Conf. Comput. Vis.* Tel Aviv, Israel: Springer, 2022, pp. 720–736.

[244] S. Sun, L. Wei, J. Xing, J. Jia, and Q. Tian, "SDDM: Score-decomposed diffusion models on manifolds for unpaired image-to-image translation," in *Proc. Int. Conf. Mach. Learn.*, 2023, pp. 33115–33134.

[245] N. Tumanyan, M. Geyer, S. Bagon, and T. Dekel, "Plug-and-play diffusion features for text-driven image-to-image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 1921–1930.

[246] S. Niklaus and F. Liu, "Context-aware synthesis for video frame interpolation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1701–1710.

[247] H. Sim, J. Oh, and M. Kim, "XVFI: EXtreme video frame interpolation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 14469–14478.

[248] Z. Liu, R. A. Yeh, X. Tang, Y. Liu, and A. Agarwala, "Video frame synthesis using deep voxel flow," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4463–4471.

[249] L. Kong et al., "IFRNet: Intermediate feature refine network for efficient frame interpolation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 1969–1978.

[250] M. Koren, K. Menda, and A. Sharma, "Frame interpolation using generative adversarial networks," Stanford Univ., Stanford, CA, USA, Tech. Rep., 2017.

[251] J. van Amersfoort et al., "Frame interpolation with multi-scale deep loss functions and generative adversarial networks," 2017, *arXiv:1711.06045*.

[252] S. Wen, W. Liu, Y. Yang, T. Huang, and Z. Zeng, "Generating realistic videos from keyframes with concatenated GANs," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 8, pp. 2337–2348, Aug. 2019.

[253] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of Wasserstein GANs," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 5769–5779.

[254] V. Voleti, A. Jolicoeur-Martineau, and C. Pal, "MCVD-masked conditional video diffusion for prediction, generation, and interpolation," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 35, 2022, pp. 23371–23385.

[255] Z. Deng, X. He, Y. Peng, X. Zhu, and L. Cheng, "MV-Diffusion: Motion-aware video diffusion model," in *Proc. 31st ACM Int. Conf. Multimedia*, Oct. 2023, pp. 7255–7263.

[256] D. Danier, F. Zhang, and D. Bull, "LDMVFI: Video frame interpolation with latent diffusion models," 2023, *arXiv:2303.09508*.

[257] J. Park, K. Ko, C. Lee, and C.-S. Kim, "BMBC: Bilateral motion estimation with bilateral cost volume for video interpolation," in *Proc. 16th Eur. Conf. Comput. Vis.*, Glasgow, U.K. Springer, Aug. 2020, pp. 109–125.

[258] D. Danier, F. Zhang, and D. Bull, "ST-MFNet: A spatio-temporal multi-flow network for frame interpolation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 3521–3531.

[259] H. Yang, P. Shi, D. Zhong, D. Pan, and Z. Ying, "Blind image quality assessment of natural distorted image based on generative adversarial networks," *IEEE Access*, vol. 7, pp. 179290–179303, 2019.

[260] J. Ma, J. Wu, L. Li, W. Dong, and X. Xie, "Active inference of GAN for no-reference image quality assessment," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, 2020, pp. 1–6.

[261] Y. Zhu, H. Ma, J. Peng, D. Liu, and Z. Xiong, "Recycling discriminator: Towards opinion-unaware image quality assessment using Wasserstein GAN," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 116–125.

[262] P. P. Ž. Babnik and V. Štruc, "DifFIQA: Face image quality assessment using denoising diffusion probabilistic models," 2023, *arXiv:2305.05768*.

[263] J. Zhang, Z. Fang, and L. Yu, "A no-reference perceptual image quality assessment database for learned image codecs," *J. Vis. Commun. Image Represent.*, vol. 88, Oct. 2022, Art. no. 103617.

[264] E. Betzalel, C. Penso, A. Navon, and E. Fetaya, "A study on the evaluation of generative models," 2022, *arXiv:2206.10935*.

[265] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 2226–2234.

[266] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 6629–6640.

[267] E. J. Nunn, P. Khadivi, and S. Samavi, "Compound Frechet inception distance for quality assessment of GAN created images," 2021, *arXiv:2106.08575*.

[268] H. Ko, D. Y. Lee, S. Cho, and A. C. Bovik, "Quality prediction on deep generative images," *IEEE Trans. Image Process.*, vol. 29, pp. 5964–5979, 2020.

[269] H. Zheng, H. Yang, J. Fu, Z.-J. Zha, and J. Luo, "Learning conditional knowledge distillation for degraded-reference image quality assessment," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10222–10231.

[270] G. Zhao, V. Magoulianitis, S. You, and C.-C.-J. Kuo, "A lightweight generalizable evaluation and enhancement framework for generative models and generated samples," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. Workshops (WACVW)*, Jan. 2024, pp. 450–459.

[271] M. Horowitz, "1.1 computing's energy problem (and what we can do about it)," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2014, pp. 10–14.

[272] J. Ballé, N. Johnston, and D. Minnen, "Integer networks for data compression with latent-variable models," in *Proc. Int. Conf. Learn. Represent.*, 2019.

[273] H. Sun, L. Yu, and J. Katto, "Learned image compression with fixed-point arithmetic," in *Proc. Picture Coding Symp. (PCS)*, Jun. 2021, pp. 1–5.

[274] D. He, Z. Yang, Y. Chen, Q. Zhang, H. Qin, and Y. Wang, "Post-training quantization for cross-platform learned image compression," 2022, *arXiv:2202.07513*.

[275] E. Koyuncu, T. Solovyev, E. Alshina, and A. Kaup, "Device interoperability for learned image compression with weights and activations quantization," in *Proc. Picture Coding Symp. (PCS)*, Dec. 2022, pp. 151–155.

[276] G.-W. Jeon, S. Yu, and J.-S. Lee, "Integer quantized learned image compression," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2023, pp. 2755–2759.

[277] E. Koyuncu, T. Solovyev, J. Sauer, E. Alshina, and A. Kaup, "Quantized decoder in learned image compression for deterministic reconstruction," 2023, *arXiv:2312.11209*.

[278] A. Gholami, S. Kim, Z. Dong, Z. Yao, M. W. Mahoney, and K. Keutzer, "A survey of quantization methods for efficient neural network inference," in *Low-Power Computer Vision*. Boca Raton, FL, USA: CRC Press, 2022, pp. 291–326.

[279] C. Jia, X. Hang, W. Liu, S. Wang, and S. Ma, "FPX-NVC: An FPGA-accelerated P-frame based neural video coding system," in *Proc. IEEE Int. Conf. Vis. Commun. Image Process. (VCIP)*, Dec. 2022, p. 1.

[280] H. Sun, L. Yu, and J. Katto, "Q-LIC: Quantizing learned image compression with channel splitting," *IEEE Trans. Circuits Syst. Video Technol.*, to be published.

[281] J. Shi, M. Lu, and Z. Ma, "Rate-distortion optimized post-training quantization for learned image compression," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 5, pp. 3082–3095, May 2024.

[282] A. Luo, H. Sun, J. Liu, and J. Katto, "Memory-efficient learned image compression with pruned hyperprior module," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Bordeaux, France, Oct. 2022, pp. 3061–3065.

[283] X. Ding et al., "ResRep: Lossless CNN pruning via decoupling remembering and forgetting," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 4490–4500.

[284] J.-H. Kim, J.-H. Choi, J. Chang, and J.-S. Lee, "Efficient deep learning-based lossy image compression via asymmetric autoencoder and pruning," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Barcelona, Spain, May 2020, pp. 2063–2067.

[285] W. Guo-Hua, J. Li, B. Li, and Y. Lu, "EVC: Towards real-time neural image compression with mask decay," in *Proc. 11th Int. Conf. Learn. Represent.*, 2022.

[286] Z. Zheng, X. Wang, X. Lin, and S. Lv, "Get the best of the three worlds: Real-time neural image compression in a non-GPU environment," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 5400–5409.

[287] S. Yin et al., "Bandwidth-efficient inference for neural image compression," 2023, *arXiv:2309.02855*.

[288] S. Yu and J.-S. Lee, "LUT-LIC: Look-up table-assisted learned image compression," in *Proc. Int. Conf. Neural Inf. Process.* Springer, 2023, pp. 430–441.

[289] C. Jia, X. Hang, S. Wang, Y. Wu, S. Ma, and W. Gao, "FPX-NIC: An FPGA-accelerated 4K ultra-high-definition neural video coding system," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 9, pp. 6385–6399, Sep. 2022.

[290] Z. Zhao, C. Jia, S. Wang, S. Ma, and J. Yang, "Learned image compression using adaptive block-wise encoding and reconstruction network," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, May 2021, pp. 1–5.

[291] H. Shao, B. Liu, Z. Li, C. Yan, Y. Sun, and T. Wang, "A high-throughput processor for GDN-based deep learning image compression," *Electronics*, vol. 12, no. 10, p. 2289, May 2023.

[292] H. Sun, Q. Yi, F. Lin, L. Yu, J. Katto, and M. Fujita, "F-LIC: FPGA-based learned image compression with a fine-grained pipeline," in *Proc. IEEE Asian Solid-State Circuits Conf. (A-SSCC)*, Nov. 2022, pp. 1–3.

[293] Z. Cheng, H. Sun, M. Takeuchi, and J. Katto, "Deep residual learning for image compression," in *Proc. CVPR Workshops*, Jun. 2019, pp. 1–5.

[294] H. Sun, Q. Yi, F. Lin, L. Yu, J. Katto, and M. Fujita, "Real-time learned image codec on FPGA," in *Proc. IEEE Int. Conf. Vis. Commun. Image Process. (VCIP)*, Dec. 2022, p. 1.

[295] S.-F. Tsai, C.-T. Li, H.-H. Chen, P.-K. Tsung, K.-Y. Chen, and L.-G. Chen, "A 1062 Mpixels/s 8192×4320p high efficiency video coding (H.265) encoder chip," in *Proc. Symp. VLSI Circuits*, Jun. 2013, pp. C188–C189.

[296] C.-T. Huang, M. Tikekar, C. Juvekar, V. Sze, and A. Chandrakasan, "A 249 Mpixel/s HEVC video-decoder chip for quad full HD applications," in *Proc. IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, Feb. 2013, pp. 162–163.

[297] D. Zhou et al., "A 4 Gpixel/s 8/10 b H.265/HEVC video decoder chip for 8K ultra HD applications," in *Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC)*, Jan. 2016, pp. 266–268.

[298] J. Liu, H. Sun, and J. Katto, "Learned image compression with mixed transformer-CNN architectures," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 14388–14397.

[299] K. Tian, Y. Guan, J. Xiang, J. Zhang, X. Han, and W. Yang, "Towards real-time neural video codec for cross-platform application using calibration information," in *Proc. 31st ACM Int. Conf. Multimedia*, Oct. 2023, pp. 7961–7970.

[300] T. Peng, G. Gao, H. Sun, F. Zhang, and D. Bull, "Accelerating learnt video codecs with gradient decay and layer-wise distillation," 2023, *arXiv:2312.02605*.

[301] S. Zhang, W. Mao, H. Shi, and Z. Wang, "A computationally efficient neural video compression accelerator based on a sparse CNN-transformer hybrid network," 2023, *arXiv:2312.10716*.

[302] H. Le et al., "MobileCodec: Neural inter-frame video compression on mobile devices," in *Proc. 13th ACM Multimedia Syst. Conf.*, Jun. 2022, pp. 324–330.

[303] T. van Rozendaal et al., "MobileNVC: Real-time 1080p neural video compression on a mobile device," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2024, pp. 4323–4333.

[304] E. Dupont, A. Golinski, M. Alizadeh, Y. W. Teh, and A. Doucet, "COIN: Compression with implicit neural representations," in *Proc. Neural Compression From Inf. Theory Appl. Workshop@ ICLR*, 2021.

[305] Y. Strümpler, J. Postels, R. Yang, L. V. Gool, and F. Tombari, "Implicit neural representations for image compression," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2022, pp. 74–91.

[306] H. Kim, M. Bauer, L. Theis, J. Richard Schwarz, and E. Dupont, "C3: High-performance and low-complexity neural compression from a single image or video," 2023, *arXiv:2312.02753*.

[307] Y. Zhang, T. van Rozendaal, J. Brehmer, M. Nagel, and T. S. Cohen, "Implicit neural video compression," in *Proc. ICLR Workshop Deep Generative Models Highly Structured Data*, 2022.

[308] T. Ladune, P. Philippe, F. Henry, G. Clare, and T. Leguay, "COOL-CHIC: Coordinate-based low complexity hierarchical image codec," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 13515–13522.

[309] T. van Rozendaal, J. Brehmer, Y. Zhang, R. Pourreza, A. Wiggers, and T. S. Cohen, "Instance-adaptive video compression: Improving neural codecs by training on the test set," 2021, *arXiv:2111.10302*.

[310] E. Dupont, H. Loya, M. Alizadeh, A. Golinski, Y. W. Teh, and A. Doucet, "COIN++: Neural compression across modalities," *Trans. Mach. Learn. Res.*, vol. 2022, pp. 1–26, Nov. 2022.

[311] C. Gomes, R. Azevedo, and C. Schroers, "Video compression with entropy-constrained neural representations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 18497–18506.

[312] B. B. Damodaran, M. Balcilar, F. Galpin, and P. Hellier, "RQAT-INR: Improved implicit neural image compression," in *Proc. Data Compress. Conf. (DCC)*, Mar. 2023, pp. 208–217.

[313] Z. Guo, G. Flamich, J. He, Z. Chen, and J. M. Hernández-Lobato, "Compression with Bayesian implicit neural representations," in *Proc. 37th Conf. Neural Inf. Process. Syst.*, 2023, pp. 1938–1956.

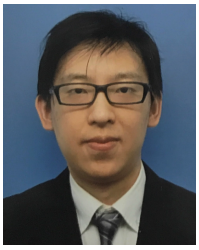
[314] G. Zhang, X. Zhang, and L. Tang, "Enhanced quantified local implicit neural representation for image compression," *IEEE Signal Process. Lett.*, vol. 30, pp. 1742–1746, 2023.

[315] C. Cremer, X. Li, and D. Duvenaud, "Inference suboptimality in variational autoencoders," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1078–1086.

[316] T. van Rozendaal, I. A. M. Huijben, and T. S. Cohen, "Overfitting for fun and profit: Instance-adaptive data compression," 2021, *arXiv:2101.08687*.

[317] M. Balcilar, B. Damodaran, and P. Hellier, "Reducing the amortization gap of entropy bottleneck in end-to-end image compression," in *Proc. Picture Coding Symp. (PCS)*, Dec. 2022, pp. 115–119.
[318] O. Jourairi, M. Balcilar, A. Lambert, and F. Schnitzler, "Improving the reconstruction quality by overfitted decoder bias in neural image compression," in *Proc. Picture Coding Symp. (PCS)*, Dec. 2022, pp. 61–65.
[319] H. Zhang, F. Cricri, H. R. Tavakoli, M. Santamaria, Y.-H. Lam, and M. M. Hannuksela, "Learn to overfit better: Finding the important parameters for learned image compression," in *Proc. Int. Conf. Vis. Commun. Image Process. (VCIP)*, Dec. 2021, pp. 1–5.
[320] M. Balcilar, B. B. Damodaran, and P. Hellier, "Reducing the mismatch between marginal and learned distributions in neural video compression," in *Proc. IEEE Int. Conf. Vis. Commun. Image Process. (VCIP)*, Dec. 2022, pp. 1–5.
[321] K. Liu et al., "Manipulation attacks on learned image compression," *IEEE Trans. Artif. Intell.*, to be published.
[322] M. Song, J. Choi, and B. Han, "A training-free defense framework for robust learned image compression," 2024, *arXiv:2401.11902*.
[323] T. Zhu et al., "Attack and defense analysis of learned image compression," 2024, *arXiv:2401.10345*.
[324] T. Chen and Z. Ma, "Toward robust neural image compression: Adversarial attack and model finetuning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 12, pp. 7842–7856, Dec. 2023.
[325] Y. Choi, M. El-Khamy, and J. Lee, "Variable rate deep image compression with a conditional autoencoder," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Seoul, South Korea, Oct. 2019, pp. 3146–3154.
[326] Z. Cui, J. Wang, S. Gao, T. Guo, Y. Feng, and B. Bai, "Asymmetric gained deep image compression with continuous rate adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 10532–10541.
[327] M. Song, J. Choi, and B. Han, "Variable-rate deep image compression through spatially-adaptive feature transform," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 2360–2369.

Zhibo Chen (Senior Member, IEEE) received the B.Sc. and Ph.D. degrees from the Department of Electrical Engineering, Tsinghua University, in 1998 and 2003, respectively. He is currently a Full Professor with the University of Science and Technology of China. He has more than 180 publications and over 100 granted patent applications. Some of his standard proposals have been adopted in MPEG/VCEG on video coding and ITU-T P.1202 on video quality assessment. His research interests focus on investigating artificial intelligence technique for advanced visual signal generation, representation, processing, and coding; and other interdisciplinary research fields. He is the Chair of IEEE Visual Signal Processing and Communications Technical Committee (VSPC-TC). He was the TPC Chair of IEEE PCS 2019, a Organization Committee Member of ICIP 2017 and ICME 2013, and a TPC Member of IEEE ISCAS and VCIP. He has served as an Associate Editor for IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY and a Guest Editor for IEEE OPEN JOURNAL OF CIRCUITS AND SYSTEMS and IEEE JOURNAL ON EMERGING AND SELECTED TOPICS IN CIRCUITS AND SYSTEMS.



Heming Sun (Member, IEEE) received the B.E. degree in electronic engineering from Shanghai Jiao Tong University, Shanghai, China, in 2011, the double M.E. degrees from Waseda University and Shanghai Jiao Tong University, in 2012 and 2014, respectively, and the Ph.D. degree from Waseda University in 2017. He was a Researcher with the NEC Central Research Laboratories, from 2017 to 2018. He was an Assistant Professor with Waseda University, from 2018 to 2023. He is currently an Associate Professor with Yokohama National University. His interests are in algorithms and VLSI architectures for image/video processing and neural networks. He is a member of the Technical Committee on Visual Signal Processing and Communications in the IEEE CAS Society. He got several awards including the IEEE Computer Society Japan Chapter Young Author Award, the IEEE VCIP Best Paper Award, and the PCS Top-Ten Best Paper Award. He is an Associate Editor of IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY.

Li Zhang (Senior Member, IEEE) received the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, in 2009. Currently, she leads the Multimedia Laboratory, ByteDance Inc., pioneering cutting-edge technologies in multimedia. With a focus on video compression, streaming, and signal processing, she holds more than 700 granted U.S. patents and has published more than 100 technical articles in book chapters, journals, and conference proceedings. Additionally, she has made more than 600 adopted standardization contributions to various standards, such as H.266/VVC, H.265/HEVC, AVS, IEEE 1857, H.264/AVC, G-PCC, and JPEG AI. Her research has been recognized with numerous awards, including the Best Paper Award at the 2022 ISCAS Visual Signal Processing and Communications Track and the Top Ten Best Paper Award at the 2021 IEEE PCS. She has also secured several first-place accolades in international challenges and received Certificates of Appreciation for her exceptional contributions to the IEEE 1857 Standard in 2013 and 2021. She has served as an editor, a software coordinator, and the chair for core experiments in standard groups. She has organized and co-chaired multiple special sessions and grand challenges at conferences. She holds the position of the Publicity Subcommittee Chair of the Technical Committee on Visual Signal Processing and Communications in the IEEE CAS Society. She is an Associate Editor of IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY.

Fan Zhang (Member, IEEE) received the B.Sc. and M.Sc. degrees from Shanghai Jiao Tong University, Shanghai, China, in 2005 and 2008, respectively, and the Ph.D. degree from the University of Bristol, Bristol, U.K., in 2012. He is currently a Senior Lecturer in visual communications with the School of Computer Science, University of Bristol. He has published over 80 academic papers, and has contributed to two books on video compression. His research interests include deep video coding, video quality assessment, perceptual video compression, and creative technology. He is a member of the Technical Committee on Visual Signal Processing and Communications in the IEEE CAS Society and the Organization Committee of PCS 2021. He has been an Associate Editor of IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, since 2021.