Generative Visual Compression: A Review

Bolin Chen*, Shanzhi Yin*, Peilin Chen*, Shiqi Wang* and Yan Ye†
* City University of Hong Kong
† Alibaba Group

***Abstract*—Artificial Intelligence Generated Content (AIGC) is leading a new technical revolution for the acquisition of digital content and impelling the progress of visual compression towards competitive performance gains and diverse functionalities over traditional codecs. This paper provides a thorough review on the recent advances of generative visual compression, illustrating great potentials and promising applications in ultra-low bitrate communication, user-specified reconstruction/filtering, and intelligent machine analysis. In particular, we review the visual data compression methodologies with deep generative models, and summarize how compact representation and high-fidelity reconstruction could be actualized via generative techniques. In addition, we generalize related generative compression technologies for machine vision and intelligent analytics. Finally, we discuss the fundamental challenges on generative visual compression techniques and envision their future research directions.**

I. Introduction

The concept "generative visual compression" was first mentioned in [1], which utilizes deep generative models (*i.e.,* Variational Auto-Encoder (VAE) [2], Generative Adversarial Network (GAN) [3] and Diffusion Model (DM) [4]) to compress the visual data in pursuit of realizing visually-pleasing reconstructions within the minimal coding costs compared with traditional image/video compression algorithms. Especially in the past two years, the cumulative integration of generative models, Contrastive Language-Image Pre-Training (CLIP) [5], Transformer [6] and other technologies has given rise to the explosion of Artificial Intelligence Generated Content (AIGC), making it more versatile in digital content creation.

Essentially, different from discriminative models predicting the decision boundary between the classes accompanied by insensitivity to outliers, deep generative models can well allow for data generation and augmentation since they focus on learning the underlying patterns or distribution from a given set of data. Similarly, visual data compression also aims to establish the relationship between an index and an image (a point) in the high dimensional image space, such that it can exploit statistical redundancy to represent data or utilize strong information prior to decompose/encode signal towards optimal rate-distortion trade-offs. As such, there are clear commonalities between the generation and compression tasks, providing great possibilities for exploring generative visual compression. Specifically, deep generative models are able to learn the compact feature distribution required in compression tasks, whilst their strong inference capabilities also facilitate the signal reconstruction from these compact distributions. Different from traditional hybrid coding frameworks such as H.264/Advanced Video Coding (AVC) [7], H.265/High Efficiency Video Coding (HEVC) [8] and H.266/ Versatile Video Coding (VVC) [9], these novel generative compression paradigms possess more compact feature representations, flexible motion estimation mechanisms, and superior signal reconstruction capabilities, which can bring promising compression performance and diverse functionalities.

This paper provides a review on generative visual compression for both human and machine visions. In particular, we generalize a wide spectrum of generative visual compression methodologies and explore the inner correlations between generation and compression. Furthermore, we introduce the technical evolution from human vision to machine vision, aiming at constructing more intelligent coding and collaborative analysis systems. Finally, we summarize basic challenges and envision future possible research for these state-of-the-art generative visual compression schemes.

II. Generative Visual Compression for Human Vision

The majority of current image/video compression research aims at improving the visually-pleasing experience for the human visual system within minimal coding bit rates. This section will review the progress of existing generative compression algorithms to clarify how well deep generative models can perform in visual compression tasks, mainly including end-to-end latent code representation, cross-modal image coding, conceptual image coding, generative coding for temporal evolution, and omni-dimensional data coding as illustrated in Fig. 1.

A. End-to-End Latent Code Representation

Benefiting from the strong inference capacity of deep generative networks, learned image/video coding algorithms can jointly optimize the entire encoder/decoder towards the rate-distortion trade-off in an end-to-end trainable manner. Ballé *et al.* [10] firstly explored the rate–distortion optimization problem within the context of VAE and proposed an end-to-end learned image compression framework, where natural image can be mapped to a latent code space via a parametric analysis transform. Following this, scale hyper-priors [11], autoregressive context model [12], Gaussian mixture model [13], channel-wise model [14], channel-spatial model [15] were successively employed to parameterize the distributions of latent codes and optimize the entropy model, achieving competitive compression performance compared to traditional codecs. In addition, the rate-distortion-complexity optimization [16], [17] was further conducted such that they can support variable

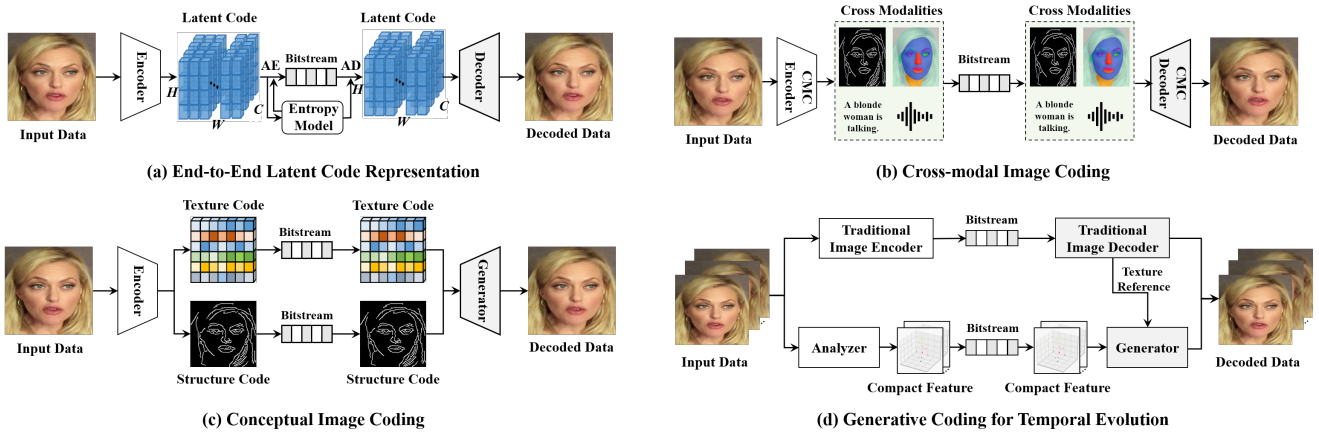

Fig. 1. Illustration of typical generative compression frameworks for human vision.

rate and flexible decoding complexity. Besides, adversarial learning [18], [19] was also introduced to improve visual quality. The recent progress of diffusion models also improved the performance of learned image compression with residual augmentation [20] and deterministic inference [21].

As for learned video compression, Habibian *et al.* [22] introduced a rate-distortion auto-encoder, and Lombardo *et al.* [23] employed a sequential VAE to exploit spatiotemporal redundancy based upon learned image codec framework. In addition, Li *et al.* [24] utilized context diversity in both temporal and spatial dimensions to enhance the neural video codec that could surpass the under-development next-generation traditional codec ECM. Furthermore, GAN-based neural video compression methods [25]–[27] have been developed for realistic synthesis and redundancy reduction.

B. Cross-modal Image Coding

Multi-modality data (*e.g.,* auditory, textual, and haptic) have been widely used in various vision tasks for robust perception and understanding. Relying on such cross-modal learning [28], visual data compression can be further developed towards high-level human-comprehensible communication. In particular, Li *et al.* [29] proposed the first cross-modal compression framework that can encode images into compact text for semantic communication. Besides, Zhang *et al.* [30] followed the philosophy of scalable coding to design a scalable cross-modality compression paradigm with different modalities. To improve the semantic fidelity, Gao *et al.* [31] developed a novel rate-distortion optimized cross-modal coding scheme using reinforcement learning. In addition, a variable-rate cross-modal compression mechanism [32] was proposed to meet the demands of changeable transmission bandwidth. Tandon *et al.* [33] presented a compression pipeline that can convert visual data into text transcript for dramatically reducing data transmission rates, whilst Lu *et al.* [34] employed different visual data types (*i.e.,* infrared and visible image pairs) to exploit the cross-modal redundancy.

C. Conceptual Image Coding

Inspired by visual decomposition tasks [35], [36], conceptual coding is proposed to disentangle natural images into a series of conceptual representations for high-efficiency compression. In details, the conceptual compression framework firstly encodes visual image data into structure information and texture code, and then achieves image decoding via a deep generative model. Following this paradigm, Chang *et al.* [37] chose to compactly encode critical structural edge and color/texture information, which can achieve promising performance over traditional image codecs at similar bitrate. On this basis, semantic prior modeling [38], consistency-contrast regularization [39] and semantic-aware visual decomposition [40] were further proposed to preserve higher coding flexibility and better visual reconstruction at extreme bitrates.

D. Generative Coding for Temporal Evolution

Current generative video coding algorithms mainly exploit strong priors or learn temporal dynamics between video frames to formulate the analysis-synthesis-based compression framework [41], thus achieving low-bandwidth communication. Specifically, the encoder employs traditional coding technique like HEVC or VVC to compress key-reference frames of video sequence and an analysis model to characterize subsequent inter frames into compact transmitted symbols, whilst the decoder can utilize the deep generative model to synthesize video frames from the decoded key-reference frames and compact temporal information.

Taking human face/body videos as examples, they possess much inherent structure and prior knowledge, such as their shape, composition, and movement, which can be easily compressed with this methodology. Based on First Order Motion Model (FOMM) [42], an end-to-end animation model, the generative video compression framework can be optimized towards ultra-low bit-rate communication [43]–[45], where video frames are characterized into a series of learned 2D keypoints. Besides, Oquab *et al.* [43] developed the first real-time generative compression system on the mobile platform via

the SPADE architecture and segmentation maps. In addition, some facial priors like 2D landmarks [46], 3D semantics [47], and compact feature representation [48] are employed as the transmitted animation symbols. Wang *et al.* [49] also proposed an ultra-low bitrate video codec via a PCA-based decomposing method, greatly allowing for the compression potential of motion representations. Regarding human body video compression, a disentangled texture-structure visual representation [50] was proposed, where human pose keypoints are leveraged as the structure code to achieve extremely low bitrate. Moreover, to increase compression performance and broaden application scenarios, various novel technologies have been applied to this generative video compression framework, such as frame interpolation [51], residual enhanced coding [52], spatial-temporal adversarial training [53], multi-view aggregation [54], multi-reference dynamic prediction [55] and feature transcoding [56]. Indeed, such low-bitrate generative video compression paradigm is entirely possible to be applied to other natural scenes [57] with oscillatory dynamics and motion prior, such as trees, flowers, candles, and clothes swaying in the wind.

E. Omni-Dimensional Data Coding

Unlike visual data from natural scenes, omni-dimensional signals like 3D point clouds, light fields, and 360-degree data can represent both static and dynamic objects/scenarios with a higher number of dimensions, thus offering the advantage of a more realistic and immersive visual experience. Indeed, this also means that the compression/transmission of such raw omni-dimensional data requires more coding bits. Thus, expanding generative visual compression algorithms into the high dimensional signal data may be one solution to achieving their compact representation and efficient reconstruction.

In particular, for the compression of point cloud geometry, Wang *et al.* [58] proposed a novel end-to-end VAE-based framework to capture compact latent features and actualize hyperprior generation. In addition, He *et al.* [59] designed an auto-encoder to preserve local density information, and Nguyen *et al.* [60] employed a deep generative model to estimate the probability distribution of voxel occupancy. As for light field compression, Jia *et al.* [61] and Liu *et al.* [62] both proposed GAN-based view synthesis-compression methods with a cascaded hierarchical coding structure and quality enhancement technique, respectively. Also, quantization-aware learning [63] and dual discrimination models [64] are utilized in light field compression tasks for promising performance. Moreover, generative compression in other high dimensional data scenes, such as stereo image [65] and 360-degree data [66], have also been explored for visual perception applications.

III. Generative Visual Compression for Machine Vision

With a considerable amount of visual data being ultimately received and processed by machine for high-level vision tasks, visual compression for machine vision has been widely studied in recent years [67]. It aims to maintain machine task performance from compressed visual data. Similar to human-oriented scenarios, many research works leverage the generative model as the backbone of machine-oriented visual compression networks and implement end-to-end optimization by combining rate, distortion, and machine performance.

In this section, we first categorized these methods based on different domains that general machine tasks are performed on. Specifically, if the machine task is performed via reconstructed pixel-level image/video, it is categorized as "Pixel Domain Analysis". On the other hand, if the machine analysis is performed directly on transformed generative feature, we categorized such methods as "Feature Domain Analysis". In addition, regardless of the requirement for machine performance, pixel-level reconstructions are available as an intermediate step for machine analysis or additional tasks in most cases. We further categorize these methods based on their arrangement for human and machine vision as "Single-branch", "Layered", "Multi-branch" or "Scalable". The illustration of those different pipelines is shown in Fig. 2.

A. Pixel Domain Analysis

1) Single-branch Approaches: The single-branch approaches of generative compression for machine vision are explored by directly feeding reconstructed visual data from learned visual data codec to machine task networks. Such works mainly focus on optimization strategies or compression network design. A study regarding end-to-end image compression for machine vision was conducted in [68] by comparing different training and tuning methods with off-the-shelf task and compression models. Le *et al.* [69] proposed a weighting schedule to balance the trade-offs among rate, distortion and task performance during training. Also, they proposed an online inference-time fine-tuning strategy to update latent features by considering the feature distortion of the task network as a proxy loss [70]. In addition, Wang *et al.* [71] designed a unified optimization framework with generalized rate-accuracy optimization and variable bit-rate coding. They also explored channel number distribution for an inverted bottleneck encoder structure [72]. To realize machine-friendly reconstruction and low bit-rate compression for underwater images, feature enhancement with prior-guided contrastive learning [73] was leveraged to improve decoder-side feature preservation and alleviate underwater degradation.

Efforts have also been made to expand the data dimension or task diversities. Yi *et al.* [74] explored the framework design and optimization such that end-to-end video compression can also achieve machine analysis, where multi-scale motion estimation and multi-frame feature fusion were proposed with task-driven optimization to trade-off between signal-level and semantic-level fidelity. To utilize multiple machine tasks, the connectors were proposed to adapt reconstructions from a primary machine task to multiple secondary tasks [75]. Similarly, Chen *et al.* [76] transferred a well-trained transformer-based image codec to various machine tasks by injecting instance-

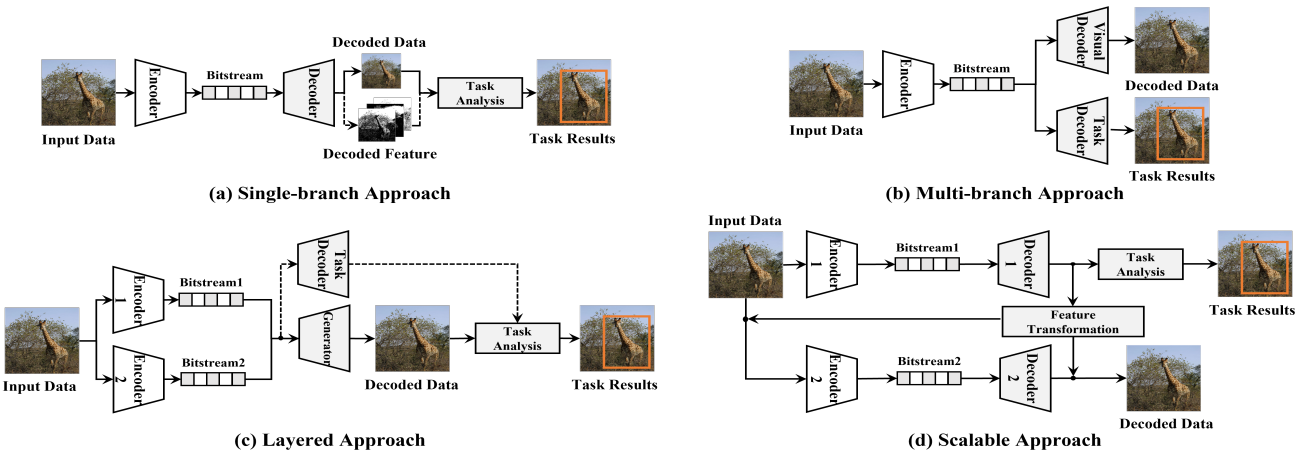

Fig. 2. Illustration of typical generative compression frameworks for machine vision. Dash lines show the optional structures.

specific prompts and task-specific prompts to the encoder and decoder.

2) Layered/Scalable Approaches: While single-branch methods are intuitive and convenient for implementation, such a naive combination of compression and machine task model lacks flexibility and hinders further performance improvement. Inspired by conceptual coding in human-oriented vision compression, there have been latest works realizing layered structure with multiple bit-streams or scalable structures with multi-level reconstructions. In the meantime, most of them take advantage of GANs to obtain ultra-low bit-rate and semantically high quality with dedicated feature designs.

Yang *et al.* [77] decomposed face images into structure code and scalable color code for sparse encoding, then recovered images with the controllable adversarial generator for both machine and human vision. Similarly, Mao *et al.* [78] proposed to disentangle face images into thumbnails and sketches for ultra-low bit-rate coding, while two-stage generative reconstructions with retrieval-guided external priors were leveraged to preserve face identity and reconstruction quality. To further utilize the strong generation ability of pretrained GAN model, a scalable GAN inversion was proposed for facial image decoding, where scalable latents were extracted by encoders for face attribute prediction and spatial feature transform of intermediate GAN features [79]. With layer-wise hierarchical semantic information from StyleGANs [80], Mao *et al.* [81] utilized hierarchical style facial features from three-layer StyleGAN2 to realize reconstructions for coarse, middle, and high-level vision tasks. Apart from facial images, underwater image features could also realize extreme compression with the aid of an edge map, saliency mask, and foreground mask [82]. An enhancement layer then compressed the residual between machine-oriented reconstruction and original input for human-oriented reconstruction.

B. Feature Domain Analysis

Instead of performing machine analysis on traditional signal-level data, there has been a trend to bypass the reconstruction procedure and directly implement machine task on generative features. The feature could be intermediate features of generative codec or specially reconstructed features in parallel to reconstructed visual data. Similar to human-oriented scenarios, some of these works use single bit-stream with varied number of decoders for different tasks, while other works generate multiple bit-streams to represent scalable information of different granularity.

1) Single-branch/Multi-branch Approaches: Torfason *et al.* [83] first explored the possibility of performing image understanding on the compressed representations by branching an inference network from latents in parallel with reconstruction decoder. Bai *et al.* [84] further extended this framework with transformer-based image analysis network and rate-distortion-accuracy optimization. To save computation for resource-constrained edge computing systems and knowledge distillation, Matsubara *et al.* [85] proposed to discard visual reconstruction and directly reconstruct features from decoder for inference with a pruned task model.

2) Layered/Scalable Approaches: To exploit the interaction between different levels of vision tasks and seek better representations for visual data, a scalable framework with multiple latents and bit-streams [86] is more common for feature domain analysis. Choi *et al.* [87] proposed to split latent representation into base layer and enhancement layer, where the former is transformed to the feature space for task back-end as well as reconstruction improvement. In addition, an adaptive partition-transmission-reconstruction-and-aggregation scheme [88] was proposed to select the optimal subset of latents of existing learning-based image codec for machine and human visions.

To further enhance the scalability of the whole network, Wang *et al.* [89] proposed to use separately encoded face feature and texture as the base layer and enhancement layer. The analysis task is based on the base layer, while reconstruction is based on both base features and residuals from the enhancement layer. Similarly, content and style of face images are separately extracted and fused for multi-

task learning [90]. Such a scalable framework has also been successfully extended to video data with a two-layer structure [91], where a semantic representation is learned to extract temporal semantic information and served for machine tasks as well as predictive coding for reconstruction. Moreover, a three-layer structure [92] was further designed with semantic, structure and texture representations to satisfy different vision requirements.

IV. Challenges and Research Directions

Generative visual compression techniques have shown great potentials in compression efficiency and versatile vision analytics for natural image and video. Nevertheless, this does not imply that such generative visual compression research has been sufficiently developed and matured. The following drawbacks and challenges still exist and need to be solved,

- **Quality measures for evaluation and optimization**: Generative visual compression algorithms are designed in the feature domain, thus not being optimized for common pixel-level quality metrics in video coding, such as Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM). In addition, although existing perceptual-based quality measures like Deep Image Structure and Texture Similarity (DISTS) [93] and Learned Perceptual Image Patch Similarity (LPIPS) [94] have been proposed for feature-domain measurement, they still lack some relevant priors and temporal learning for generative assessment. As a result, it is imperative to use appropriate perceptual assessment measures like [95] in order to precisely evaluate the performance effectiveness and optimize the algorithm design of generative compression tasks.
- **Robustness & generalization capability**: Limited by the upper capabilities of deep generative models, the current reconstruction of generative compression algorithms is sometimes faced with some visual distortions and artifacts, resulting in an unpleasant visual quality of experience. Such instability in reconstruction quality damages the feasibility of practical applications to a certain extent. Therefore, it is important to explore more mature and robust technologies based on existing algorithms to improve the generalization ability of generative visual compression models.
- **Task-dependent compression & communication**: These existing generative visual compression algorithms are mostly designed for specialized natural scenes, such as face and human body data with strong priors, so they are not task-independent and scenario-irrelevant. In real-scene applications, there is a high probability that visual data mixed with various objects/scenes will be encountered, which these existing algorithms cannot handle as well as traditional hybrid coding frameworks. As such, it is important to design a unified and universal generative compression algorithm for task-independent and scene-collaborated communication.
- **Standardization & deployment**: In October 2023, the Joint Video Experts Team (JVET) of the ISO/IEC SC 29 and ITU-T SG16 has established a new Ad Hoc Group to conduct investigations [96], [97] on generative face video compression regarding software implementation, test conditions, coordinated experimentation, interoperability study, model lightweight and other aspects. Taking this as a starting point, it is hopeful that the standardization and deployment of other generative visual compression algorithms can be explored in the near future. In addition, to further optimize product design and user experience, including cost, performance, and power, hardware-software co-design of generative visual compression algorithm should be taken into account.

V. Conclusions

This paper has made a comprehensive review on generative visual compression for both human and machine visions. Thanks to the strong inference capability of deep generative models, relevant visual compression for image and video data could be optimized and developed towards high coding efficiency, realistic signal reconstructions, and intelligent task analysis. As such, huge volumes of visual data around the world can be efficiently compressed via generative compression, whilst new applications and requirements in the post-AIGC era will further accelerate the research progress of generative visual compression techniques.

References

[1] Shibani Santurkar et al., "Generative compression," in *Picture Coding Symposium*, 2018, pp. 258–262.

[2] Diederik P Kingma and Max Welling, "Auto-encoding variational bayes," in *Proc. Int. Conf. Learn. Represent.*, 2014, p. 14.

[3] Ian Goodfellow et al., "Generative adversarial nets," *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014.

[4] Jonathan Ho et al., "Denoising diffusion probabilistic models," *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 6840–6851, 2020.

[5] Alec Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.

[6] Ashish Vaswani et al., "Attention is all you need," *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[7] Thomas Wiegand et al., "Overview of the H.264/AVC video coding standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 7, pp. 560–576, 2003.

[8] Gary J Sullivan et al., "Overview of the High Efficiency Video Coding (HEVC) standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1649–1668, 2012.

[9] Benjamin Bross et al., "Overview of the Versatile Video Coding (VVC) standard and its applications," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 10, 2021.

[10] Johannes Ballé et al., "End-to-end optimized image compression," in *Proc. Int. Conf. Learn. Represent.*, 2017.

[11] Johannes Ballé et al., "Variational image compression with a scale hyperprior," in *Proc. Int. Conf. Learn. Represent.*, 2018.

[12] David Minnen et al., "Joint autoregressive and hierarchical priors for learned image compression," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, vol. 31.

[13] Zhengxue Cheng et al., "Learned image compression with discretized gaussian mixture likelihoods and attention modules," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, June 2020.

[14] David Minnen et al., "Channel-wise autoregressive entropy models for learned image compression," in *Proc. IEEE Int. Conf. Image Process.* IEEE, 2020, pp. 3339–3343.

[15] Dailan He et al., "ELIC: Efficient learned image compression with unevenly grouped space-channel contextual adaptive coding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 5718–5727.

[16] Zhichen Zhang et al., "ELFIC: A learning-based flexible image codec with rate-distortion-complexity optimization," in *Proc. ACM Int. Conf. Multimedia*, 2023, pp. 9252–9261.

[17] Fei Yang et al., "Slimmable compressive autoencoders for practical neural image compression," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[18] Eirikur Agustsson et al., "Generative adversarial networks for extreme learned image compression," in *Proc. IEEE Int. Conf. Comput. Vis.*, October 2019.

[19] Fabian Mentzer et al., "High-fidelity generative image compression," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 11913–11924.

[20] Noor Fathima Goose et al., "Neural image compression with a diffusion-based decoder," *arXiv preprint arXiv:2301.05489*, 2023.

[21] Ruihan Yang and Stephan Mandt, "Lossy image compression with conditional diffusion models," *arXiv preprint arXiv:2209.06950*, 2022.

[22] Amirhossein Habibian et al., "Video compression with rate-distortion autoencoders," in *Proc. IEEE Int. Conf. Comput. Vis.*, October 2019.

[23] Salvator Lombardo et al., "Deep generative video compression," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, vol. 32.

[24] Jiahao Li et al., "Neural video compression with diverse contexts," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, June 2023, pp. 22616–22626.

[25] Fabian Mentzer et al., "Neural video compression using GANs for detail synthesis and propagation," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 562–578.

[26] Tiesong Zhao et al., "Learning-based video coding with joint deep compression and enhancement," in *Proc. ACM Int. Conf. Multimedia*, 2022, p. 3045–3054.

[27] Bowen Liu et al., "Deep learning in latent space for video prediction and compression," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, June 2021, pp. 701–710.

[28] Tadas Baltrušaitis et al., "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, 2019.

[29] Jiguo Li et al., "Cross modal compression: Towards human-comprehensible semantic compression," in *Proc. ACM Int. Conf. Multimedia*, 2021, p. 4230–4238.

[30] Pingping Zhang et al., "Rethinking semantic image compression: Scalable representation with cross-modality transfer," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 8, pp. 4441–4445, 2023.

[31] Junlong Gao et al., "Rate-distortion optimization for cross modal compression," in *Data Compression Conference*, 2023, pp. 218–227.

[32] Junlong Gao et al., "Cross modal compression with variable rate prompt," *IEEE Trans. Multimedia*, pp. 1–13, 2023.

[33] Pulkit Tandon et al., "Txt2vid: Ultra-low bitrate compression of talking-head videos via text," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 1, pp. 107–118, 2023.

[34] Guo Lu et al., "Learning based multi-modality image and video compression," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, June 2022, pp. 6083–6092.

[35] Junho Jeon et al., "Intrinsic image decomposition using structure-texture separation and surface normals," in *European Conference Computer Vision*. Springer, 2014, pp. 218–233.

[36] Youngjung Kim et al., "Structure-texture image decomposition using deep variational priors," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2692–2704, 2018.

[37] Jianhui Chang et al., "Layered conceptual image compression via deep semantic synthesis," in *Proc. IEEE Int. Conf. Image Process.*, 2019, pp. 694–698.

[38] Jianhui Chang et al., "Conceptual compression via deep structure and texture synthesis," *IEEE Trans. Image Process.*, vol. 31, pp. 2809–2823, 2022.

[39] Jianhui Chang et al., "Consistency-contrast learning for conceptual coding," in *Proc. ACM Int. Conf. Multimedia*, 2022, pp. 2681–2690.

[40] Jianhui Chang et al., "Semantic-aware visual decomposition for image coding," *Int. J. Comput. Vis.*, pp. 1–23, 2023.

[41] Bolin Chen et al., "Generative face video coding techniques and standardization efforts: A review," *arXiv preprint arXiv:2311.02649*, 2023.

[42] Aliaksandr Siarohin et al., "First order motion model for image animation," *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019.

[43] Maxime Oquab et al., "Low bandwidth video-chat compression using deep generative models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshop*, 2021.

[44] Goluck Konuko et al., "Ultra-low bitrate video conferencing using deep image animation," in *IEEE Int. Conf. Acoust. Speech Signal Process.*, 2021.

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

[46] Dahu Feng et al., "A generative compression framework for low bandwidth video conference," in *Proc. IEEE Int. Conf. Multimedia Expo. Workshops*. IEEE, 2021.

[47] Bolin Chen et al., "Interactive face video coding: A generative compression framework," *arXiv preprint arXiv:2302.09919*, 2023.

[48] Bolin Chen et al., "Beyond keypoint coding: Temporal evolution inference with compact feature representation for talking face video compression," in *Data Compression Conference*, 2022.

[49] Ruofan Wang et al., "Extreme generative human-oriented video coding via motion representation compression," in *IEEE International Symposium on Circuits and Systems*, 2023, pp. 1–5.

[50] Ruofan Wang et al., "Disentangled visual representations for extreme human body video compression," in *Proc. IEEE Int. Conf. Multimedia Expo.*, 2022, pp. 1–6.

[51] Madhav Agarwal et al., "Compressing video calls using synthetic talking heads," in *British Machine Vision Conference*, 2023.

[52] Goluck Konuko et al., "Predictive coding for animation-based video compression," in *Proc. IEEE Int. Conf. Image Process.*, 2023.

[53] Bolin Chen et al., "Compact temporal trajectory representation for talking face video compression," *IEEE Trans. Circuits Syst. Video Technol.*, 2023.

[54] Anna Volokitin et al., "Neural face video compression using multiple views," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022.

[55] Zhao Wang et al., "Dynamic multi-reference generative prediction for face video compression," in *Proc. IEEE Int. Conf. Image Process.*, 2022, pp. 896–900.

[56] Shanzhi Yin et al., "Enabling translatability of generative face video coding: A unified face feature transcoding framework," in *Data Compression Conference*, 2024.

[57] Zhengqi Li et al., "Generative image dynamics," *arXiv preprint arXiv:2309.07906*, 2023.

[58] Jianqiang Wang et al., "Lossy point cloud geometry compression via end-to-end learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 12, pp. 4909–4923, 2021.

[59] Yun He et al., "Density-preserving deep point cloud compression," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 2333–2342.

[60] Dat Thanh Nguyen et al., "Lossless coding of point cloud geometry using a deep generative model," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 12, pp. 4617–4629, 2021.

[61] Chuanmin Jia et al., "Light field image compression using generative adversarial network-based view synthesis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 9, no. 1, pp. 177–189, 2018.

[62] Deyang Liu et al., "View synthesis-based light field image compression using a generative adversarial network," *Information Sciences*, vol. 545, pp. 118–131, 2021.

[63] Xiaoran Jiang et al., "An untrained neural network prior for light field compression," *IEEE Trans. Image Process.*, vol. 31, 2022.

[64] Nader Bakir et al., "Light field image coding using dual discriminator generative adversarial network and vvc temporal scalability," in *Proc. IEEE Int. Conf. Multimedia Expo.* IEEE, 2020, pp. 1–6.

[65] Jerry Liu et al., "Dsic: Deep stereo image compression," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 3136–3145.

[66] Mu Li et al., "End-to-end optimized 360° image compression," *IEEE Trans. Image Process.*, vol. 31, pp. 6267–6281, 2022.

[67] Lingyu Duan et al., "Video coding for machines: A paradigm of collaborative compression and intelligent analytics," *IEEE Trans. Image Process.*, vol. 29, pp. 8680–8695, 2020.

[68] Lahiru D. Chamain et al., "End-to-end optimized image compression for machines, a study," in *Data Compression Conference*, 2021.

[69] Nam Le et al., "Image coding for machines: an end-to-end learned approach," in *IEEE Int. Conf. Acoust. Speech Signal Process.*, 2021, pp. 1590–1594.

[70] Nam Le et al., "Learned image coding for machines: A content-adaptive approach," in *Proc. IEEE Int. Conf. Multimedia Expo.*, 2021.

[71] Shurun Wang et al., "Deep image compression toward machine vision: A unified optimization framework," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 6, pp. 2979–2989, 2023.
[72] Shurun Wang et al., "End-to-end compression towards machine vision: Network architecture design and optimization," *IEEE open j. circuits syst.*, vol. 2, pp. 675–685, 2021.
[73] Zhengkai Fang et al., "Prior-guided contrastive image compression for underwater machine vision," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 6, pp. 2950–2961, 2023.
[74] Xiaokai Yi et al., "Task-driven video compression for humans and machines: Framework design and optimization," *IEEE Trans. Multimedia*, vol. 25, pp. 8091–8102, 2023.
[75] Lahiru D Chamain et al., "End-to-end optimized image compression for multiple machine tasks," *arXiv preprint arXiv:2103.04178*, 2021.
[76] Yi-Hsin Chen et al., "Transtic: Transferring transformer-based image compression from human perception to machine perception," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2023, pp. 23240–23250.
[77] Shuai Yang et al., "Towards coding for human and machine vision: Scalable face image coding," *IEEE Trans. Multimedia*, vol. 23, pp. 2957–2971, 2021.
[78] Yudong Mao et al., "Peering into the sketch: Ultra-low bitrate face compression for joint human and machine perception," in *Proc. ACM Int. Conf. Multimedia*, 2023, p. 2564–2572.
[79] Wenhan Yang et al., "Facial image compression via neural image manifold compression," *IEEE Trans. Circuits Syst. Video Technol.*, pp. 1–1, 2023.
[80] Tero Karras et al., "Analyzing and improving the image quality of stylegan," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020.
[81] Qi Mao et al., "Scalable face image coding via stylegan prior: Toward compression for human-machine collaborative vision," *IEEE Trans. Image Process.*, vol. 33, pp. 408–422, 2024.
[82] Zhengkai Fang et al., "Priors guided extreme underwater image compression for machine vision and human vision," *IEEE J. Ocean. Eng.*, vol. 48, no. 3, pp. 888–902, 2023.
[83] Robert Torfason et al., "Towards image understanding from deep compression without decoding," *arXiv preprint arXiv:1803.06131*, 2018.
[84] Yuanchao Bai et al., "Towards end-to-end image compression and analysis with transformers," in *Proc. AAAI Conf. Artif. Intell.*, 2022.
[85] Yoshitomo Matsubara et al., "Supervised compression for resource-constrained edge computing systems," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 2685–2695.
[86] Yuefeng Zhang et al., "Machine perception-driven image compression: A layered generative approach," *arXiv preprint arXiv:2304.06896*, 2023.
[87] Hyomin Choi et al., "Scalable image coding for humans and machines," *IEEE Trans. Image Process.*, vol. 31, pp. 2739–2754, 2022.
[88] Lei Liu et al., "Icmh-net: Neural image compression towards both machine vision and human vision," in *Proc. ACM Int. Conf. Multimedia*, 2023, p. 8047–8056.
[89] Shurun Wang et al., "Towards analysis-friendly face representation with scalable feature and texture compression," *IEEE Trans. Multimedia*, vol. 24, pp. 3169–3181, 2022.
[90] Yuefeng Zhang et al., "Analysis on compressed domain: A multi-task learning approach," in *Data Compression Conference*, 2022, p. 494.
[91] Zhimeng Huang et al., "Hmfvc: A human-machine friendly video compression scheme," *IEEE Trans. Circuits Syst. Video Technol.*, pp. 1–1, 2022.
[92] Hongbin Lin et al., "DeepSVC: Deep scalable video coding for both machine and human vision," in *Proc. ACM Int. Conf. Multimedia*, 2023, p. 9205–9214.
[93] Keyan Ding et al., "Image quality assessment: Unifying structure and texture similarity," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2020.
[94] Richard Zhang et al., "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2018.
[95] Yixuan Li et al., "Perceptual quality assessment of face video compression: A benchmark and an effective method," *arXiv preprint arXiv:2304.07056*, 2023.
[96] Yan Ye et al., "On VVC-assisted ultra-low rate generative face video coding," *MPEG ISO/IEC JTC 1/SC 29/WG 2 doc. no. m64987*, October 2023.
[97] Bolin Chen et al., "AHG16: Proposed common software tools and testing conditions for generative face video compression," *The Joint Video Experts Team of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29, doc. no. JVET-AG0042*, January 2024.