APSIPA Transactions on Signal and Information Processing, 2024, 13, e502

Original Paper

Human-Machine Collaborative Image and Video Compression: A Survey

Huanyang Li[1,2], Xinfeng Zhang[1*], Shiqi Wang[3], Shanshe Wang[4] and Jingshan Pan[5]

[1] *University of Chinese Academy of Sciences, Beijing, China*
[2] *Pengcheng Laboratory, Shenzhen, China*
[3] *School of Computer Science, City University of Hong Kong, Hong Kong*
[4] *Information Technology R&D Innovation Center of Peking University, China*
[5] *Shandong Computer Science Center (National Supercomputer Center in Jinan), China*

ABSTRACT

Traditional image and video compression methods are designed to maintain the quality of human visual perception, which makes it necessary to reconstruct the image or video before machine analysis. Compression methods oriented towards machine vision tasks make it possible to use the bit stream directly for machine vision tasks, but it is difficult for them to decode high quality images. To bridge the gap between machine vision tasks and signal-level representation, researchers present plenty of the human-machine collaborative compression methods. In order to provide researchers with a comprehensive understanding of this field and promote the development of image and video compression, we present this survey. In this work, we give a problem definition and explore the relationship and application scenarios of different methods. In addition, we provide a comparative analysis of existing methods on compression and machine vision tasks performance. Finally, we provide a discussion of several directions that are most promising for future research.

*Corresponding author: Xinfeng Zhang, xfzhang@ucas.ac.cn

Received 31 July 2024; revised 26 August 2024; accepted 10 September 2024
ISSN 2048-7703; DOI 10.1561/116.20240052

Keywords: Image and video compression, human perceptual quality, machine analysis.

1 Introduction

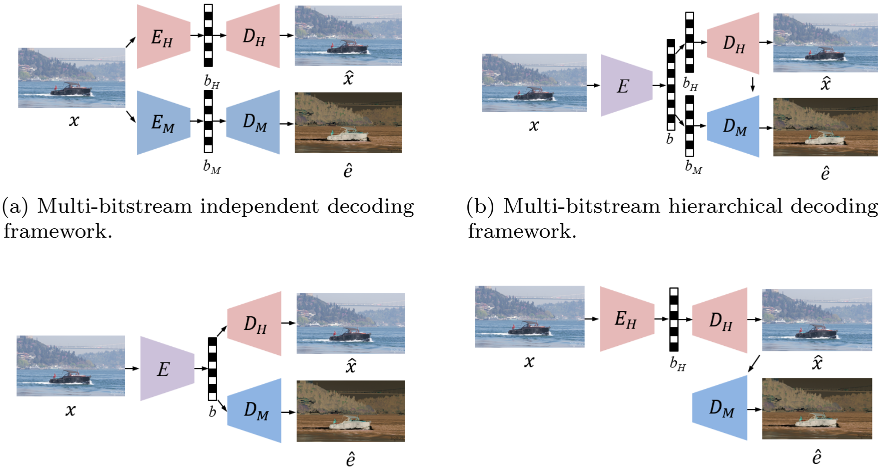
In recent years, the data volume of images and videos has experienced explosive growth due to the development of Internet. A large amount of images and videos are produced, stored, transmitted and processed. Thus, image and video compression technology plays an essential role to reduce the bandwidth and space for data transmission and storage while maintaining the visual quality. The traditional aim of image and video compression is to optimize the quality of human visual perception at a certain bit rate, making the quality of compressed image and video close to that of original one. To achieve this goal, a series of traditional compression techniques for images and videos are proposed, such as discrete cosine transform (DCT), motion compensation, inter-frame prediction, quantization and entropy coding. These technologies have made great progress in the past few decades and have formed a series of standards and specifications, such as JPEG [180], JPEG2000 [133], AVC [191], HEVC [20], VVC [21], AV1 [32], AVS3 [209]. These standards collectively have driven the evolution of image and video storage, transmission, and analysis, adequately addressing the human requirements for the quality of images and videos in the digital age. In addition, with the development of deep learning, some efficient compression methods based on neural networks have been proposed [15, 143, 144, 157, 169, 2, 53, 52, 35, 7, 218, 31, 101, 60, 12, 79, 153, 36, 102, 168, 109, 119, 107, 208, 99, 23, 150, 6, 199, 88, 42, 172, 173, 181, 151, 94, 77, 145, 175, 3, 163, 210, 95, 192, 87, 58, 17, 201, 96, 1, 134, 135, 97, 63, 137, 71, 78, 138, 198, 40, 43, 22, 187, 118, 212, 108, 156, 110, 194]. These methods also primarily focus on the quality of human visual reconstruction. When dealing with machine vision tasks, people have to decode the image or video before machine analysis, which hampers the compression process from efficiently fulfilling the requirements of machine vision systems.

The rapid development of artificial intelligence also leads to increasingly widespread applications of machine vision across various domains: deep learning models are employed to tackle complex tasks such as image and video classification [74, 73, 126, 141, 193, 84, 47, 140, 139, 203, 91, 65, 92, 25, 127, 61, 154, 130, 66], object detection [216, 67, 69, 106, 57, 159, 152, 44, 112, 113, 183, 124, 70, 81, 116, 100, 146, 115, 215, 83], and object segmentation [149, 68, 103, 13, 19, 28, 182, 204, 147, 29, 179], which means that machines have become an important recipients and processors of images and videos. However, decoding high-quality images and videos before machine analysis brings significant computational costs, while decoding low-quality images and videos may results in poor feature extraction, thus reducing analysis performance.

To meet the diverse requirements of machine vision, relevant image and video compression standards for machine vision are continuously being developed and refined such as CDVS [49] and CDVA [50], which aim to generate compact descriptors to support specific tasks like image and video retrieval and visual search. In addition to standards, the academic community also propose a series of related image and video feature compression methods [8, 33, 34, 76, 161, 166, 167, 206] to improve the analysis efficiency of machine vision. However, the compressed features are unable to reconstruct images or videos to meet human visual demands. Considering the necessity of human-machine collaborative compression, the international organization for standardization established relevant standards for image and video compression technologies. For instance, Moving Picture Experts Group Video Coding for Machines (MPEG VCM) [51] aims to provide efficient video compression and feature extraction techniques to support video data processing and machine vision tasks. Besides, JPEG AI standard [11] is proposed to facilitate the efficient distribution and machine consumption of images. It emphasizes the utilization of advanced image compression methods based on DNN to surpass the compression efficacy of conventional methods. In addition to the above standards, numerous technologies have been proposed to address human-machine collaborative image and video compression issues. As shown in Figure 1, these methods can be categorized into four types based on the components of the compressed information and their decoding approaches: multi-bitstream independent decoding (MBID) [49, 125, 18, 162, 148, 90, 24, 120], multi-bitstream hierarchical decoding (MBHD) [9, 111, 185, 131, 184, 80, 205, 30, 5, 72, 190, 104, 165, 56, 121, 54, 202, 196, 122, 37, 213, 55, 105, 38, 195, 14] , single-bitstream multi-head decoding (SBMD) [11, 123, 26, 176], and single-bitstream analysis after reconstruction (SBAR)[132, 186, 62, 59]. These methods not only ensure compression efficiency, but also take into account the needs of both human and machine vision tasks.

Based on the above works, several surveys have summarized the work in the compression field. Some surveys summarize the innovative work in the field of learning based image and video compression [129, 89, 214]. In [214], Zhang *et al.* summarize and compare perceptually optimized video compression methods. Some surveys take into account of the gap between machine analysis and signal-level reconstruction. Ma *et al.* [128] provide an overview of joint feature and texture representation frameworks. Dong and Pan [48] summarize the connections between compression and machine vision tasks.

These works provide summaries and outlooks on the field of image and video compression. In recent years, a series of human-machine collaborative encoding methods have been proposed, which can satisfy both high-level and low-level tasks at lower bit rates. On one hand, these methods address the issue that compression techniques oriented towards human vision are inefficient



(a) Multi-bitstream independent decoding framework.

(b) Multi-bitstream hierarchical decoding framework.

(c) Single-bitstream multi-head decoding framework.

(d) Single-bitstream analysis after reconstruction framework.

Figure 1: Different human-machine collaborative image compression frameworks. E represents the encoder, D represents the decoder, E_H and D_H represent the human visual compression codec, E_M and D_M represent the machine vision task codec, b represents the bitstream, x represents the original image, $\hat{x}$ represents the reconstructed image, and $\hat{e}$ represents the features used for machine vision tasks.

for machine vision tasks. On the other hand, they solve the problem that machine-oriented compression methods have difficulty in reconstructing signal-level representations to a great extent. Therefore, this paper aims to provide a comprehensive overview of human-machine collaborative image and video compression. The main contributions of this paper can be summarized as follows:

- We provide a comprehensive review on image and video compression methods that cater to both human visual perception and machine analysis requirements, analyzing the motivations and principles of these methods.
- We analyze the performances of the reviewed human-machine collaborative methods on commonly used benchmarks.
- We identify some potential challenges and directions in the human-machine collaborative image and video compression domain.

We have made every effort to collect the vast majority of papers related to this field. The rest of this overview is organized as follows: Section 2 defines the problem of human-machine collaborative image and video compression, and introduces relevant metrics for human visual perception and machine analysis. Section 3 introduces the categories of human-machine collaborative image compression and provides analysis of the methods. Section 4 classifies and discusses human-machine collaborative video compression methods. Section 5 provides performance comparisons of these methods. Section 6 discusses remaining challenges and potential research directions and concludes the survey.

2 Foundations of Human-Machine Collaborative Image and Video Compression

2.1 Problem Definition

For human-machine collaborative image and video compression, the most important problem is how to achieve a balance between the quality of human visual reconstruction and the efficiency of machine analysis. Given an image or video $\boldsymbol{x}$, compression frameworks designed for human recipients primarily aim to minimize the compression bitrate while maintaining high visual quality. Consequently, the optimization objective focused on human visual perception can be articulated as follows:

$$L_{human} = \min(D(\boldsymbol{x}, \hat{\boldsymbol{x}}) + \lambda R), \tag{1}$$

where R denotes the amount of bits in the bitstream that needs to be transmitted. The bitstream includes compressed image or video data. Sometimes it also contains network information such as the network parameters of Implicit Neural Representation (INR). λ is a balancing parameter, and D measures the distortion between the original image or video $\boldsymbol{x}$ and the reconstructed image or video $\hat{\boldsymbol{x}}$ obtained through compression.

Beyond assessing the reconstruction quality and compression bitrate of the image or video, it is a new trend to consider the requirements of machine vision tasks. For a given set of N machine vision tasks with their corresponding labels $Y = \{Y_1, Y_2, \ldots, Y_N\}$, we denote $F = \{F_1, F_2, \ldots, F_N\}$ as the features extracted from x for these tasks and denote $\hat{Y_i}$ as the predicted outcome for task i. We define $L_i(\hat{Y_i}, Y_i)$ as the loss for task i in relation to the features $\hat{F_i}$ and labels $\hat{Y_i}$ derived from the decoded image or video. Considering the varying importance of different tasks, we introduce weighting parameters to define the optimization objective for machine vision tasks as follows:

$$L_{machine} = \min\left(\lambda R + \textstyle\sum_{i=1}^{N} \omega_i L_i\right), \tag{2}$$

where ω_i is weight parameters utilized to balance the significance of each task. R denotes the bitrate of image or video features. By integrating the optimization objectives for human visual reconstruction and machine analysis, we formulate a comprehensive optimization objective function for human-machine collaborative image and video compression, aiming to minimize the bitrate costs and the loss of human and machine vision tasks:

$$L = L_{human} + L_{machine} = \min(\omega_0 D(\boldsymbol{x}, \hat{\boldsymbol{x}}) + \lambda R + \textstyle\sum_{i=1}^{N} \omega_i L_i), \tag{3}$$

2.2 Compression Performance Metric

We summarize two primary categories of metrics used to evaluate the performance of compression algorithms: human visual metrics and machine analysis metrics, which ensures a comprehensive evaluation of the impact of compression on both human viewers and machine vision tasks.

2.2.1 Human Visual Metric

Human visual metrics are designed to measure the quality of a compressed image or video from the perspective of human viewers. These metrics are crucial for ensuring that compressed content remains visually pleasing. The primary metrics include Peak Signal-to-Noise Ratio (PSNR) [75], Structural Similarity Index (SSIM) [188] and Multi-Scale Structural Similarity (MSSSIM) [189].

2.2.2 Machine Analysis Metrics

For machine analysis, metrics are designed to evaluate the performance of machine vision analysis algorithms on some specific image and video tasks such as classification, object detection, and object segmentation. For classification task, the widely used metric is classification accuracy. For object detection, precision, recall, F1-Score [200], and Intersection over Union (IoU) [155] are employed to measure both the accuracy and the overlap of predicted object boundaries against the ground truth. For segmentation, IoU, Dice Coefficient [158], and Pixel Accuracy are pivotal in measuring the accuracy of boundary delineation and the similarity between predicted and true segmentation.

3 Human-Machine Collaborative Image Compression

In order to obtain compact representations that can support both pixel-level reconstruction and semantic analysis, numerous methods have been proposed. As we mentioned in the first section, these methods can be categorized into four categories: MBID, MBHD, SBMD, SBAR. These methods will be discussed in detail in subsequent sections. Table 1 provides a comprehensive summary of them.

Table 1: An overview of human-machine collaborative image compression methods in literature. MBID, MBHD, SBMD and SBAR respectively represent multi-bitstream independent decoding, multi-bitstream hierarchical decoding, single-bitstream multi-head decoding, and single-bitstream analysis after reconstruction. The ✓indicates that the method aims to reconstruct and analyze facial images.

Category	Author	Presented Task	Core Method	Facial Image Specific
MBID	[111]	recognition	TFQI-based joint bit allocation	
	[177]	classification	cross-layer context model + ROI	
	[27]	segmentation	semantic feature enhancement	
	[24]	detection	slimmable compressive encoder	
	[120]	detection segmentation	gate module+knowledge distillation	
MBHD	[9]	face detection	feature map + CNN	✓
	[185]	face recognition	feature & texture representation	✓
	[131]	facial identity recognition, facial attribute prediction	StyleGAN prior + layer-wise scalable entropy transformer	✓
	[184]	face verification	feature & texture + residual	✓
	[80]	facial landmark detection	edge map + GAN	✓
	[205]	segmentation	GAN+hyperprior model	✓
	[30]	detection	instance segmentation map + signal feature	
	[5]	image search	semantic segmentation map + residual	
	[72]	semantic enhancement	semantic segmentation + enhancement	
	[190]	classification	task feature+residual	
	[104]	classification	residual enhance + GAN	
	[165]	detection	object separation + parameter share	
	[56]	segmentation, pose estimation	customized group mask + group-independent transform	
	[121]	classification	pyramid of multiple subbands	
	[54]	face recognition	Canny edge color sketch	✓
	[202]	detection, segmentation	structural representation+VGG	
	[196]	detection	depth-constrained encoder	
	[122]	classification, detection, segmentation	hyperprior network + predictor module	
	[37]	detection	latent space transform	
	[213]	classification, segmentation	reconstruction semantic feature fusion	
	[55]	detection, segmentation	structural edges + feature + prior	
	[105]	classification	semantics-based ROI mask + generation module	
	[38]	detection, segmentation	ask-dependent latent space transform	
	[195]	detection	mask multilayer fusion	
	[14]	classification	lightweight image encoder+ViT	
SBMD	[123]	classification	general feature extraction + feature-analytic classifier	
	[26]	classification, detection, segmentation	prompt generator + Transformer	
	[176]	classification, segmentation	feature-maps	
SBAR	[132]	face recognition	sketches thumbnails + retrieved guidance	
	[186]	detection	inverted bottleneck structure encoder	
	[62]	detection, segmentation, facial landmark detection	content-adaptive diffusion model	
	[59]	image caption, detection	feature distance + importance-weighted pixel distance	

3.1 Multi-bitstream independent decoding

In addition to the bitstream used for image reconstruction, MBID methods introduce an additional independent bitstream by extracting features and compressing them to support high-level tasks. Some methods use local image descriptors for machine vision tasks, such as the Scale-Invariant Feature Transform (SIFT) proposed by Lowe [125] and the Speeded Up Robust Features (SURF) introduced by Bay *et al.* [18]. Other approaches utilize global image descriptors to summarize high-level image properties for advanced analysis. Sivic and Zisserman [162] address large-scale image search using the bag-of-visual-words (BOV), while Perronnin *et al.* [148] focuse on compressing Fisher vectors to reduce memory usage and accelerate retrieval, aiming to supplant the bag-of-visual-words technique. Additionally, Jégou *et al.* [90] design a simplified version of the Fisher kernel representation to tackle the challenge of image search on a very large scale. A representative work is Compact Descriptors for Visual Search (CDVS) [49] , which extracts and compresses local and global features into an independent bitstream to support efficient mobile visual search task (Figure 2).

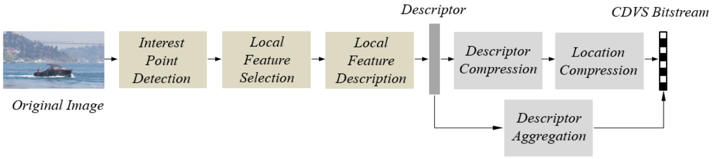


Figure 2: CDVS standard normative blocks. Global and local features are extracted from the image and compressed.

In addition to methods that extract features using traditional computer vision techniques, several learning-based MBID methods have also been proposed in recent years. To support various machine analysis needs across different task scenarios, Liu *et al.* [120] develop a method to optimize machine vision tasks in the compressed domain. This work could avoid complex decoding processes and directly performing machine vision tasks on compressed representations. Gating modules are used to select features and transformation modules to process images. Besides, it employs knowledge distillation to improve accuracy and support multitask processing. Cao *et al.* [24] introduce an adjustable multitask image compression method that balances human and machine vision needs on resource-constrained devices. By designing CNN compressors with different channel numbers for machine vision and human vision, this method not only ensures the reconstruction quality but also improves the performance of machine analysis.

3.2 Multi-bitstream hierarchical decoding

MBID methods adds extra machine vision task stream to the traditional human visual reconstruction stream, which increases the storage burdens. In order to avoid this issue, researchers developed the human-machine collaborative image compression methods that supports hierarchical decoding of the stream. Subsets of the compressed stream are utilized to perform machine vision tasks. They can be integrated with the remaining streams to reconstruct images. Among various machine vision tasks, facial tasks are of great significance because of their widespread application in daily life. Researchers propose some methods to process facial images for facial analysis tasks specifically. We will discuss these methods separately. Besides, most methods are designed for general image analysis tasks. Some methods incorporate semantic information, while others employ adaptive frameworks for machine vision tasks. We will discuss them in turn.

3.2.1 Compression methods for facial task

An early work [9] directly extracts features from the HEVC encoded bitstream. This method significantly reduces processing time by skipping traditional decoding steps such as dequantization and inverse transformation. It employs squared patches and convolutional networks for face detection, achieving efficient detection speed and accuracy. It's particularly suited for processing static images or I-frames of encoded videos.

Similarly, several studies design various methods to extract facial textures for machine vision tasks related to face recognition. Wang *et al.* [185] introduce a scalable facial image compression approach that includes a basic layer for feature compression and an enhancement layer for texture reconstruction. This method leverages deep learning models for feature extraction and texture information reconstruction. Mao *et al.* [131] utilize a StyleGAN-based approach to encode face image in scalable style, allowing flexible control over image quality and semantic information through multi-layer encoding. This method provides superior visual performance at extremely low bitrates, and is suitable for low-resolution facial image applications. In addition to directly extracting texture features, other methods improve face reconstruction quality by introducing additional information. Wang *et al.* [184] introduce a ramework contains basic and enhancement layers. The base layer extract feature for machine vision tasks and coarse reconstrunction. The enhancement layer take the residuals between coarse reconstrunction image and original image as inputs to enhance the texture information. The enhanced residuals are utilized to decode the high quality image in conjunction with the coarse reconstruction image. Fang *et al.* [54] proposed a face image compression framework. The original image is converted into a designed color sparse sketch using image-to-image transformation. This transformation helps to reduce the redundancy in the image. The sketch can be used for machine vision tasks and reconstruction. The multiscale discriminator of the framework is designed to enhance the detail information. Hu *et al.* [80] transform images into edge maps and key reference pixels, optimizing feature representation compactness and reducing required encoding bits. This method is able to meet the requirements of machine vision tasks such as facial landmark detection, it also can reconstruct high-quality image. Yang *et al.* [205] combine generative models and deep learning techniques to achieve ultra-low bitrate facial image compression. It compresses and transmits highly compact feature vectors, which are transformable for machine analysis. This framework mainly supports face segmentation.

3.2.2 Semantic Information Based Compression Methods

Facial image analysis tasks are just one part of machine vision tasks, most of the methods aims to meet the machine analysis requirements for general images, not just facial images. Some researchers designed various frameworks to utilize semantic segmentation for human-machine collaborative image compression. On one hand, the semantic segmentation maps can be used to enhance image quality. On the other hand, this kind of methods can support machine vision tasks such as object segmentation at a lower bit rate. For example, Akbari *et al.* [5] propose a framework for image compression that utilizes deep learning and semantic segmentation. The input image and its corresponding segmentation map are used to generate a compact representation to obtain a coarse reconstruction of the image. The residuals of coarse reconstruction are transmitted to enhance the visual quality. Based on this work [5], Hoang *et al.* [72] introduce a method enhancing image reconstruction quality through semantic segmentation. It utilizes specially structured neural networks to map deformation semantic back to the original distribution of semantic segmentation, enhancing the performance of image compression. In 2021, Chen *et al.* [27] propose an end-to-end mutually enhancing network for image compression and semantic segmentation. This method uses traditional image compression algorithms to compress the input image into a low-bit-rate encoded image. Its semantic segmentation module employs advanced semantic segmentation networks to generate a semantic segmentation map. The enhancement module utilizes the semantic information extracted from the semantic segmentation map to improve the image quality. In addition, Feng *et al.* [56] explore an image compression method based on irregular group decoupling and customized semantic partitions for efficient image reconstruction. This approach supports object detection and instance segmentation. It also allows the encryption of specific image parts to enhance data security and compression efficiency. In addition to directly using the semantic segmentation map for compression, some works extract advanced semantic information for machine analysis and enhanced reconstruction quality. Tu *et al.* [177] introduce a cross-layer context model to reduce redundancy and improve compression efficiency. This method takes higher-layer features as cross-layer priors. The compression mechanism is applied only to the ROIs. The generated scalable bitstream can be partially decoded for specific machine vision tasks or fully decoded for human viewing. Chen *et al.* [30] extract gray-scale profile to satisfy the demind of machine analysis such as classification, detection, and segmentation. Gray-scale profile along with low-level signal features are combined to generate the low quality image. The high quality image is reconstructed using the low quality image and the residual map. Zhang *et al.* [213] utilize a layered generative approach for machine perception-driven image compression. The method consists of a learning-based layered compression model and a multi-task analysis network. The learning-based layered compression model includes an encoder, a decoder, and a probability estimation model. The encoder encodes the input image into reconstruction part and semantic part. A fusion module is used for reconstruction. The multi-task analysis network is designed to perform machine vision tasks on the compressed representation such as classification and segmentation.

3.2.3 Other Compression Methods

In addition to the two categories mentioned above, there are some methods that make innovations in hierarchical codec framework. Wu *et al.* [196] propose a task adaptive network to support image compression for both human vision and machine vision tasks. The training process of this network is guided by a teacher network. The quantized latent representation of latent representation can be used to reconstruct different levels of images through multi-scale decoders. Similarly, Wang *et al.* [190] propose a two-stage approach which contains a feature domain analysis network and a preview image generation network. It encodes the input images into quantized analysis-oriented feature maps, which can be directly used by the machine analysis algorithm without reconstructing the RGB images. Feature residual and feature maps are then combined to reconstruct a high-quality image. Choi and Bajić *et al.* [38] present a scalable multi-task image compression method. It split the latent space into base part and enhancement part. The base part is used for machine vision tasks and the full latent space is used for reconstruction. The content of the transmission depends on the needs of downstream tasks.

In addition, some other methods make advantage of different deep learning base models to improve image quality and machine analysis accuracy. Bai *et al.* [14] encodes images into discrete representations and uses the Transformers for decoding and analysis, including dedicated classifiers and reconstructors. A key advantage of this approach is leveraging Transformers' global information processing capabilities. Lei *et al.* [104] propose a progressive deep image compression (DIC) scheme for image classification and reconstruction. They utilizes semantics analysis module classifies the input image. Class activation mapping is used to generates a semantic importance map of latent vector. Generative Adversarial Networks (GAN) is adopted to improve perceptual quality by matching the reconstructed image to the input image in the statistical domain.

3.3 Single-bitstream multi-head decoding

The previously discussed method uses multi-stream hierarchical decoding to meet multitask requirements. Besides, some single-stream methods transform the entire stream and utilize different task decoders to address human and machine vision tasks. Torfason *et al.* [176] explore a method that use the compressed representations for machine inference. Instead of decoding the compressed representation into RGB space, the authors integrate the encoders and decoders of DNN-based compression methods with architectures for image understanding. This approach reduces computational cost and allows for inference on the compressed representations. Liu *et al.* [123] propose a versatile framework that integrates image compression task and image classification task. The goal is to extract a fully-shared latent representation that supports both compression and classification. The framework extract features and utilize classifier to get compact and general shared latent representations. Similarly, Chen *et al.* [26] proposed a method to use a trained Transformer-based image codec for machine inference without fine-tuning the codec. The method utilizes prompting techniques to achieve this transfer. The instance prompt is fed into the encoder and the task prompt is fed into the decoder. The decoded image is made suitable for machine vision tasks such as object detection.

3.4 Single-bitstream analysis after reconstruction

The aforementioned SBMD frameworks meet machine vision task requirements with multi-decoders. In addition to these frameworks, there are methods that introduce machine vision task related image information to improve machine analysis performance after image reconstruction.

Mao *et al.* [132] utilize learned facial image compression methods based on external prior knowledge. It encodes facial images into sketches and thumbnails, and combine them to reconstruction, which improves the quality and analytic performance of reconstruction facial images. Wang *et al.* [186] propose an end-to-end deep image compression framework for machine vision tasks, which utilizes inverted bottleneck structure to optimize channel distribution.

This structure uses compact semantic feature representation to optimize rate-accuracy performance. Guo *et al.* [62] employ content-adaptive and diffusion techniques for image feature compression. This method allows flexible switching between different perceptual quality standards at extremely low bit rates. It utilizes contrastive learning and pseudo-label techniques significantly enhances the perceptual quality and encoding performance of images.

Coding Optimization-based Research: In addition to the methods based on coding networks, there are also approaches based on coding optimization that enhance human-machine collaborative image compression efficiency by implementing adjustable quantization techniques and other optimization schemes. Li *et al.* [111] design the texture feature quality index to guide compression. In order to improve both reconstruction quality and recognition accuracy, they combine the HEVC/H.265 standard for texture encoding with scalar quantization and deep feature entropy coding. Lei *et al.* [105] design an adaptive image compression method. It selects regions of interest (ROI) based on their semantic importance. The encoder and decoder calculate a ROI gain matrix and a ROI inverse gain matrix to control the quantization accuracy of different latent vector elements. Gao *et al.* [59] design a multitask image compression method, introducing an optimization strategy based on semantic metrics. By adjusting the compression network's quantization steps and distortion measures through bit allocation and semantic metrics, it reduces distortion while preserving semantic information. The reconstructed images are suited for various machine analysis tasks.

4 Human-Machine Collaborative Video Compression

Compared with images, there is a temporal correlation between video frames. This makes human-machine collaborative image compression methods inadequate to meet the compression requirements of videos. In order to solve this problem, researchers developed several human-machine collaborative video compression methods. Since human-machine collaborative video encoding methods using the SBMD framework have been found yet, the existing human-machine collaborative video compression methods can be classified into three types similarly: MBID, MBHD, and SBAR. Table 2 provides a comprehensive summary of them.

4.1 Multi-bitstream independent decoding

The CDVA standard is a representative of this category of methods. Duan *et al.* [50] offer a compact and efficient representation of video feature descriptors. It reduces redundancy through keyframe detection and extracting potent deep learning features using convolutional neural networks (CNNs) combined with Nested Invariance Pooling (NIP) technology. This standard optimizes video

Table 2: An overview of human-machine collaborative video compression methods in literature. MBID, MBHD, and SBMD respectively represent multi-bitstream independent decoding, multi-bitstream hierarchical decoding, and single-bitstream multi-head decoding.

Category	Author	Presented Task	Core Method
MBID	[50]	Video Retrieval	Feature extrAction + CDVS + CNN
	[211]	Video Retrieval	Rate-accuracy optimization + affine motion compensation
	[10]	Class Identification, Object Recognition	Comprising Multiple autoencoders
MBHD	[197]	Action Recognition	Conditional deep generation network
	[82]	Action Recognition	Semantic information + feature Laddering Framework
	[114]	Object Detection	Conditional semantic compression + interlayer frame prediction
	[64]	Object Detection	End-to-end learnable video codec + conditional coding
	[39]	Object Detection	Conventional + DNN video compression
	[85]	Action Recognition	Learned semantic representation + end-to-end optimize
	[93]	Object Detection, Pose Estimation, Action Recognition, Object Segmentation	Static Object characteristic + dynamic motion clue
	[170]	Action Recognition, Multiple Object Tracking, Object Segmentation	Traditional codec + DNN
	[171]	Action Recognition, Multiple Object Tracking, Object Segmentation	Semantic-Mining-then-Compensation + masked image modeling
	[4]	Object Detection	Cuboidal feature descriptor
SBMD	[207]	Action Recognition	Task-driven optimization
	[160]	Action Recognition, Object Detection, Object Tracking, Object Segmentation	Temporal context + cross-domain motion

structure and reduces computational complexity. Similarly, Zhang *et al.* [211] utilize feature-based affine motion compensation technology to optimize video quality and feature retrieval capabilities. This approach merges video streams and feature data into a bitstream with robust visual retrieval capabilities, which can support local feature descriptors such as SIFT, SURF, and CDVS. Antonio *et al.* [10] propose a visual objects compression method for smart surveillance applications. Several autoencoders are adopted to produce a compact latent representation of a specific object class.

4.2 Multi-bitstream hierarchical decoding

This kind of method employs hierarchical compression strategies to dynamically adapt to different decoding requirements. Some hierarchical methods analyze intra-frame information and inter-frame relationships within the stream to support analysis and reconstruction. Choi and Bajić *et al.* [39] propose a two-layer scalable video compression framework, which combines conventional and learning-based video compression techniques. The base layer contains the information related to object detection, and the enhancement layer is designed for high-quality reconstruction. Hadizadeh and Bajić [64] introduce a scalable video compression framework that consists of a base layer and an enhancement layer. In the base layer, the video frames are encoded into a compressed base bitstream. The decoded base frames are utilized by a computer vision model for video analysis, specifically object detection. The enhancement layer compresses the input frames conditionally to generate a compressed bitstream. The enhancement layer's decoder then reconstructs the output frames for human viewing. In addition, some other hierarchical compression methods improve the performance of video compression by embedding deep semantic information into the compression process. In 2022, Huang *et al.* [82] proposed a visual compression framework that consists of three layers. The basic layer compress the semantic information for machine vision tasks. The enrichment layer focuses on pixel-level information and is used for tasks such as semantic segmentation and human parsing. Key frames are compressed separately. The visual layer use the decoded content from basic layer and enrichment layer to reconstruct high quality video, which reduces the transmission burden. Besides, Huang *et al.* [85] proposed a jointly end to end video compression framework. It extracts semantic information between temporal neighboring frames, which can support both signal reconstruction and machine analysis. In 2023, Lin *et al.* [114] proposed a scalable video compression framework. It consists of three main components: compact representations, scalable bitstream, and video compression. This method extracts compact representations from videos, including semantic features, structure features, and texture features. These representations are then compressed into a scalable bitstream. A conditional semantic compression module is designed to reduce spatial-temporal redundancy of the semantic feature. An interlayer frame prediction module models the interlayer correlation and predicts video frames using the semantic feature. Jin *et al.* [93] introduced an innovative semantic video compression method that incorporates static and dynamic visual clues into a structured bitstream to support machine vision tasks. By generating a Semantic Structured Bitstream (SSB), this method significantly reduces the cost and complexity of video decompression while enabling direct processing by machine algorithms. Tian *et al.* [171] employed an autoencoder network that aims to compress videos while preserving the semantic information in an unsupervised manner. The method utilizes a mask autoencoder to learn a compact representation of the video frames. It's trained with a combination of semantic loss and nonsemantic suppression loss. In 2024, Tian *et al.* [170] proposed a compression framework that aims to integrate traditional video codecs with neural network-based models, which preserves the semantic content during compression. The authors emphasize the importance of task-decoupled design principles, scalable compression, label-free learning schemes, and effective semantic priors in an AI-task-oriented video compression system. The proposed framework incorporates these principles and aims to provide a versatile compression system that supports diverse tasks.

Furthermore, some hierarchical compression methods utilize innovative compression strategies related to the characteristics of the video. Xia *et al.* [197] present a joint compression framework for surveillance scenes, which utilizes a learnable sparse motion pattern to guide the generation of video frames through a deep generative model. This approach reduces the total coding cost of both features and videos. Ahmmed *et al.* [4] present a collaborative video compression method that utilizes cuboidal partitioning. This technique divides video frames into multiple cuboids to extract and encode features, which significantly reduces the bit requirements and computational complexity. This strategy could meet the requirements of both reconstruction and machine vision tasks such as object detection. Ikusan and R. Dai [86] introduce an intermediate feature compression framework, which consists of several components including feature extraction, feature selection, rate-distortion optimization, and video encoding. CNN is used for feature extraction, and hierarchical clustering technology is utilized to select the most relevant features. The selected features are reconstructed for different machine vision tasks.

4.3 Single-bitstream multi-head decoding

Similar to image compression method, some human-machine collaborative video methods are designed with multiple decoding units to cater to various tasks.

Yi *et al.* [207] introduce a task-driven video compression framework that enhances video quality and compression efficiency through optimized multi-scale motion estimation and multi-frame feature fusion. Moreover, the framework utilizes multitask learning approaches to optimize the encoding process, aiming to balance signal and semantic fidelity. Sheng *et al.* [160] present VNVC, a multifunctional neural video compression framework that supports various video tasks. The framework includes video reconstruction, enhancement, and analysis module, using a single bitstream with multiple decoding modules. It decodes videos partially into intermediate features that are directly available for downstream tasks, thereby reducing decoding complexity and enhancing task performance.

5 Comparative Analysis of Techniques

In this section, we discuss the details of the performance evaluation of the human-machine collaborative compression methods. First, we introduce some commonly used reconstruction benchmark databases. Next, we provide a detailed discussion of the reconstruction performance of various methods. Then, we introduce various machine vision benchmark databases that are used for the evaluation purpose. Finally, we compare the machine analysis performance of the methods. Some papers provide official open-sourced page and code links. We summarize the links in Table 3.

Table 3: The open-sourced code links of human-machine collaborative image and video compression methods.

Author	Code Link
Hu *et al.* [80]	https://williamyang1991.github.io/projects/VCM-Face/
Akbari *et al.* [5]	https://github.com/makbari7/DSSLIC
Fang *et al.* [55]	https://global.iflytek.com/
Torfason *et al.* [176]	https://github.com/DrSleep/tensorflow-deeplab-resnet
Gao *et al.* [59]	https://github.com/chansongoal/semantic_image_compression
Xia *et al.* [197]	https://lists.aau.at/mailman/listinfo/mpe-vcm
Lin *et al.* [114]	https://github.com/LHB116/DeepSVC.
Huang *et al.* [85]	https://github.com/ZhihaoHu/PyTorchVideoCompression
Yi *et al.* [207]	https://mic.tongji.edu.cn.

5.1 Image and video compression methods preformance

5.1.1 Human Oriented Compression Performance

We first compare the performance of compression frameworks in the image domain. When the recipient is human, the compression method aims to address the rate-distortion optimization problem. There are several commonly used image compression databases such as Kodak [98], CLIC2020 [174], and ImageNet [46]. Kodak database consists of 24 high-quality images originally provided by Eastman Kodak Company. These images are typically used to test the performance of various image compression techniques because they include a wide range of real-world scenes and are known for their high resolution and quality.

ImageNet database is a computer vision dataset created by Professor Li of Stanford University. The database contains 14,197,122 images and 21,841 Synset indexes. Synset is a node in the WordNet hierarchy, which is a set of synonyms. The ImageNet dataset has always been a benchmark for evaluating the performance of image classification algorithms. Object information and bounding boxes are also provided.

CLIC2020 database is a part of an annual image compression competition. The database includes a variety of images that test the abilities of compression algorithms under real-world conditions. This database contains images with varying resolutions and lighting conditions, which is comprehensive to assess the performance of learning based image compression methods.

Some papers present compression performance of their methods on Kodak [105, 195, 56, 27, 72, 5, 38, 30, 195, 72]. The rate distortion (RD) curves of these methods are shown in Figure 3.

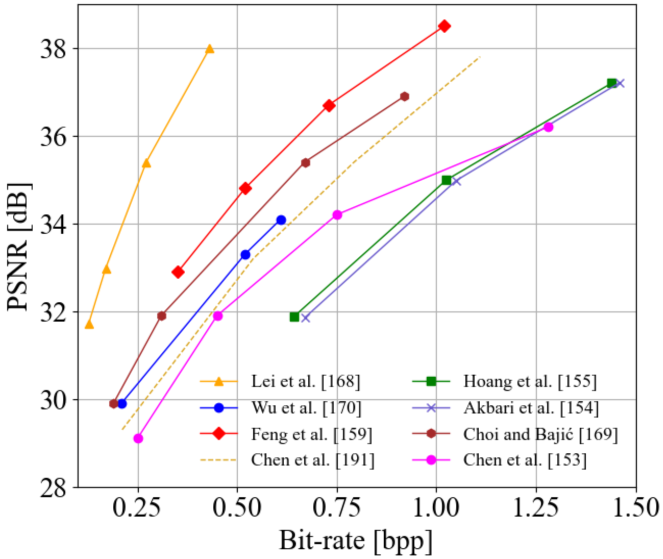


Figure 3: Rate-distortion performance evaluations for the latest compression methods on Kodak. Methods belonging to the multi-bitstream independent decoding category are represented with dashed lines, while multi-bitstream hierarchical decoding methods are represented with solid lines.

Similarly, there are some commonly used datasets in the field of video compression such as HEVC [142] and UVG [136]. HEVC Test Sequences are a set of carefully selected video clips specifically designed to evaluate and optimize the performance of HEVC codecs. These sequences cover a wide range of resolutions, from low resolution to ultra-high definition (such as 4K), and include diverse scene types and content, such as motion scenes, natural landscapes, and computer-generated imagery. HEVC Test Sequences played a crucial role during the standardization process of HEVC and serve as essential resources for developing and validating new video compression technologies.

The UVG dataset is a widely used resource in the fields of video compression and quality assessment, which is released by the Ultra Video Group at Tampere University in Finland. This dataset provides high-quality test material for evaluating video encoding, decoding, and quality assessment techniques. The UVG dataset features 4K resolution (3840x2160) video clips with diverse content types, including motion scenes, natural landscapes, computer-generated imagery, and animations. The clips are recorded at a high frame rate (120 fps), allowing researchers to assess codec performance and video quality under high frame rate conditions.

Considering that most paper use HEVC class B as the compression test set, we use RD curves to compare the compression performance of these methods in Figure 4.

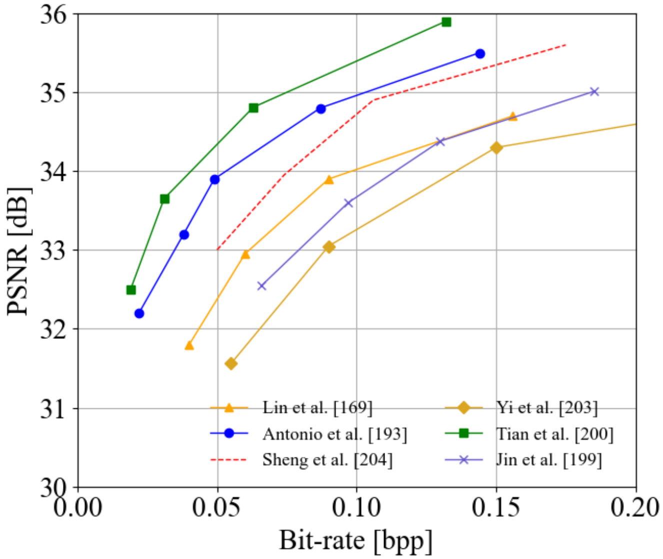


Figure 4: Rate-distortion performance evaluations for the latest human-machine collaborative video compression methods on HEVC class B. Methods belonging to the single-bitstream multi-head decoding category are represented with dashed lines, while multi-bitstream hierarchical decoding methods are represented with solid lines.

5.1.2 Machine Analysis Performance

In addition to support lossy reconstruction of images, human-machine collaborative compression methods also support one or more machine vision tasks such as classification, object detection, and object segmentation. Commonly used databases include: Caltech101 [16], Pascal VOC 2012 [178], COCO [117], LFW [217], Cityscapes [41], UCF101 [164], and MOT17 [45].

Caltech101 contains images from 101 different object categories. Each category includes about 40 to 800 images. Categories range from various types of animals, objects, and scenes. This database is commonly used for image classification and object recognition tasks.

Pascal VOC 2012 is part of the PASCAL Visual Object Classes Challenge. It's a widely used database for object detection, image segmentation, and classification. It includes images from 20 categories such as animals, vehicles, and household items, with annotations for object detection, segmentation, and classification.

COCO is a large-scale database for object detection, segmentation, and captioning. Most of the images are taken from everyday scenes and natural environments. The database includes label information for object segmentation, object localization, and image captioning.

LFW focuses on face recognition and consists of JPEG images collected over the internet. The person name is labeled. LFW is used for studies in automatic face recognition.

Cityscapes provides a large database of urban street scenes for semantic urban scene understanding. It contains a diverse set of stereo video sequences recorded in street scenes from 50 different cities, with annotations for semantic urban scene understanding tasks such as segmentation.

UCF101 is a widely used action recognition database. Released by the University of Central Florida, this database contains 13,320 video clips which belong to 101 action categories, such as sports activities, daily actions, and human-object interactions. The videos are collected from YouTube and offer a diverse range of scenes and camera motions, providing a comprehensive benchmark for evaluating action recognition algorithms.

MOT17 is a benchmark dataset widely used for evaluating multi-object tracking algorithms in computer vision. Released as part of the MOTChallenge, it includes a diverse set of video sequences recorded in various challenging real-world scenarios, such as busy streets and public spaces, with multiple pedestrians and vehicles. Each video is annotated with precise bounding boxes and unique identifiers for each object, providing ground truth for tracking performance evaluation.

We select a number of recent human-machine collaborative image and video compression algorithms from different categories and compare their performance on image classification, object detection, and object segmentation. Tables 4, 5 and 6 displays the machine vision task performance of selected human-machine collaborative image and video compression methods in the corresponding databases. The "-" indicates that the bitrate information is not provided in original paper. Blank space indicates that the method does not support the task.

Table 4: Image machine vision task performance of facial analysis tasks. The symbol "—" means the bitrate information is not given in paper or the task performance is not related to bitrate.

Author	bitrate	face recognition Acc.	face detection recall	NME	face seg Acc.
Li *et al.* [111]	0.07	0.99 (LFW)			
Alvar *et al.* [9]	—		0.98 (LFW)		
Wang *et al.* [185]	0.1	0.98 (LFW)			
Mao *et al.* [131]	0.01	0.75 (CelebA-HQ)			
Wang *et al.* [184]	0.003	0.99 (LFW)			
Hu *et al.* [80]	0.225			3.5 (VGGFace2)	
Yang *et al.* [205]	0.004				0.83 (FFHQ-Aging)
Fang *et al.* [54]	0.05	0.95 (LFW)			

Table 5: Image machine vision task performance of classification, detection and segmentation task. The symbol "—" means the bitrate information is not given in paper or the task performance is not related to bitrate.

Author	bitrate	Classifacation Acc.	Detection mAP	Segmentation mAP	Seg IoU
Tu *et al.* [177]	0.002	0.92 (CUB-200-2011)			
Chen *et al.* [27]	—				0.728 (Cityscapes)
Cao *et al.* [24]	0.2		0.54 (COCO2014)		
Liu *et al.* [120]	0.125		0.48 (PASCAL VOC 07)		0.70 (Cityscapes)
Chen *et al.* [30]	—			0.428 (COCO2017)	
Akbari *et al.* [5]	—				
Hoang *et al.* [72]	—				
Wang *et al.* [190]	0.5	0.91 (ImageNet)			
Lei *et al.* [104]	0.2	0.8 (ImageNet)			
Sun *et al.* [165]		0.75 (COCO2014)			
Feng *et al.* [56]	0.45			0.362 (COCO 2017)	
Liu *et al.* [121]	0.15	0.98 (CUB200-2011)			
Yan *et al.* [202]	—	0.9085 (CUB)			
Wu *et al.* [196]	0.1		0.745 (COCO)		
Liu *et al.* [122]	0.4	0.71 (ILSVRC2012)	0.73 (VOC2012)	0.365 (COCO)	
Choi and Bajić *et al.* [37]	0.24		0.55 (COCO 2014)		
Zhang *et al.* [213]	0.175		0.8841 (CelebAMask-HQ)		0.58 (CelebAMask-HQ)
Fang *et al.* [55]	0.035		0.512 (SUIM)		0.545 (SUIM)
Lei *et al.* [105]	0.2	0.83 (ImageNet)			
Choi and Bajić *et al.* [38]	0.15		0.39 (COCO 2014)	0.362 (COCO 2014)	
Wu *et al.* [195]	0.2		0.74 (VOC2007)		
Bai *et al.* [14]	0.2	0.73 (ImageNet)			
Liu *et al.* [123]	0.18	0.77 (Caltec)			
Chen *et al.* [26]	0.2	0.75 (ImageNet)	0.39 (COCO2017)	0.361 (COCO2017)	
Torfason *et al.* [176]	0.098	0.5582 (ILSVRC2012)			0.5578 (ILSVRC2012)
Mao *et al.* [132]	0.0281	0.5538 (CelebAHQ)			
Wang *et al.* [186]	0.4		0.48 (COCO2017)		
Guo *et al.* [62]	0.153		0.376 (COCO 2017)	0.335 (COCO 2017)	
Gao *et al.* [59]	0.2		0.48 (COCO 2014)		

Table 6: Video machine vision task performance of classification, detection and segmentation task. The symbol "—" means the bitrate information is not given in paper or the task performance is not related to bitrate.

Author	bitrate (bpp)	Detection mAP	Action Recognition Acc.	MOTA	J&F
Xia *et al.* [197]	0.0052		0.746 (PKU-MMD)		
Huang *et al.* [82]	0.013		0.751 (PKU-MMD)		
Lin *et al.* [114]	0.05	0.738 (ImageNet VID)			
Hadizadeh and Bajić [64]	0.04	0.617 (HEVC Class B)			
Huang *et al.* [85]	—		0.9905 (UCF-101)		
Jin *et al.* [93]	0.11	0.39 (COCO2014)	0.85 (UCF-101)		
Tian *et al.* [170]	0.03		0.8939 (UCF-101)	0.74 (MOT17)	0.83 (DAVIS2017)
Tian *et al.* [171]	0.02		0.75 (UCF-101)	0.75 (MOT17)	0.74 (DAVIS2017)
Yi *et al.* [207]	0.1		0.853 (UCF101)		
Sheng *et al.* [160]	0.1	0.723 (ImageNet VID)	0.504 (UCF101)	0.534 (MOT17)	0.551 (DAVIS2017)

6 Conclusion and Future Directions

The majority of compressed images and videos are ultimately intended for human viewing or machine processing. To meet the requirements of human visual perception and machine analysis, significant strides have been made in the realm of human-machine collaborative compression. This paper presents and synthesizes recent advancements in human-machine collaborative image and video compression methods. These methods not only ensure visual quality for humans but also boost utility for machine vision tasks such as classification, object detection, and object segmentation. We categorize them into 4 categories. In addition, we summarized comparative evaluations of some advanced methods in various tasks. However, the existing methods primarily focus on conventional visual tasks for images and videos. It might be challenging for them to accomplish machine vision tasks such as video summarization, object counting, and zero-shot classification. Furthermore, the utilization of large models and prior knowledge might be a potential direction. Based on the current development of image and video compression techniques, we think the following content may be promising topic for further improving performance of human and machine collaborative compression methods.

- Large models with extensive prior knowledge may be able to further enhance the performance of compression algorithms, particularly in managing complex or low-bitrate images and videos. These models could aid in predicting essential content, optimizing bit allocation, and minimizing visual redundancy.

- Cross-model compression may be more suitable to the two kinds of recipients, which can simultaneously hand image/video, audio, point cloud and text, which corresponds to visual information and semantic information. This might boost the compression efficiency and enhances functionalities in applications such as video captioning and multimedia searches. Besides, it could improve the efficiency of intelligent analysis and automated decision making, which makes contribution to the development of industrial applications such as autonomous driving and robotics.

- Furthermore, the combination of handcrafted feature representation and deep learning based representation may be able to provide a promising balance between compression performance and generalization for machine vision tasks.

Acknowledgements

This work was supported by the National Natural Science Foundation under Grant 62071449 and U20A20184, the Fundamental Research Funds for the Central Universities (E2ET1104) and the major project of PCL under Grant PCL2023A08.

References

[1] A. H. Abbas, A. Arab, and J. Harbi, "Image compression using principal component analysis", in *Mustansiriyah Journal of Science*, Vol. 29, No. 2, 2018.

[2] E. Agustsson, F. Mentzer, M. Tschannen, L. Cavigelli, R. Timofte, L. Benini, and L. V. Gool, "Soft-to-Hard Vector Quantization for End-to-End Learning Compressible Representations", in *Proceedings of Advances in Neural Information Processing Systems*, 2017, 1141–51.

[3] E. Agustsson, M. Tschannen, F. Mentzer, R. Timofte, and L. V. Gool, "Generative adversarial networks for extreme learned image compression", in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, 221–31.

[4] A. Ahmmed, M. Paul, M. Murshed, and D. Taubman, "Human-machine collaborative video coding through cuboidal partitioning", in *2021 IEEE International Conference on Image Processing*, 2021, 2074–8.

[5] M. Akbari, J. Liang, and J. Han, "DSSLIC: Deep semantic segmentation-based layered image compression", in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019, 2042–6.

[6] P. Akyazi and T. Ebrahimi, "Learning-based image compression using convolutional autoencoder and wavelet decomposition", in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

[7] D. Alexandre, C.-P. Chang, W.-H. Peng, and H.-M. Hang, "An autoencoder-based learned image compressor: Description of challenge proposal by nctu", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, 2539–42.

[8] S. R. Alvar and I. V. Bajić, "Multi-task learning with compressible features for collaborative intelligence", in *Proceedings of the IEEE International Conference on Image Processing*, 2019, 1705–9.

[9] S. R. Alvar, H. Choi, and I. V. Bajic, "Can you tell a face from a HEVC bitstream?", in *2018 IEEE Conference on Multimedia Information Processing and Retrieval*, 2018, 257–61.

[10] R. Antonio, S. Faria, L. M. Tavora, A. Navarro, and P. Assuncao, "Learning-based compression of visual objects for smart surveillance", in *2022 Eleventh International Conference on Image Processing Theory, Tools and Applications (IPTA)*, 2022, 1–6.

[11] J. Ascenso, E. Alshina, and T. Ebrahimi, "The jpeg ai standard: Providing efficient human and machine visual data consumption", in *IEEE Multimedia*, Vol. 30, No. 1, 2023, 100–11.

[12] S. Ayzik and S. Avidan, "Deep image compression using decoder side information", in *Proceedings of Computer Vision-ECCV 2020: 16th European Conference*, 2020, 699–714.

[13] M. Bai and R. Urtasun, "Deep watershed transform for instance segmentation", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, 5221–9.

[14] Y. Bai, X. Yang, X. Liu, J. Jiang, Y. Wang, X. Ji, and W. Gao, "Towards end-to-end image compression and analysis with transformers", in *Proceedings of the AAAI conference on artificial intelligence*, Vol. 36, No. 1, 2022, 104–12.

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

[16] M. Bansal, M. Kumar, M. Sachdeva, and A. Mittal, "Transfer learning for image classification using VGG19: Caltech-101 image data set", in *Journal of ambient intelligence and humanized computing*, 2023, 1–12.

[17] D. Báscones, C. González, and D. Mozos, "Hyperspectral image compression using vector quantization, PCA and JPEG2000", in *Remote Sensing*, Vol. 10, No. 6, 2018, 907.

[18] H. Bay, T. Tuytelaars, and L. V. Gool, "SURF: Speeded up robust features", in *Computer Vision—ECCV*, 2006, 404–17.

[19] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, "Yolact: Real-time instance segmentation", in *Proceedings of the IEEE international conference on computer vision*, 2019, 9157–66.

[20] F. Bossen, B. Bross, K. Suhring, and D. Flynn, "HEVC complexity and implementation analysis", in *IEEE Transactions on circuits and Systems for Video Technology*, Vol. 22, No. 12, 2012, 1685–96.

[21] B. Bross, Y.-K. Wang, Y. Ye, S. Liu, J. Chen, G. J. Sullivan, and J.-R. Ohm, "Overview of the versatile video coding (VVC) standard and its applications", in *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 31, No. 10, 2021, 3736–64.

[22] A. Buades, B. Coll, and J.-M. Morel, "A non-local algorithm for image denoising", in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Vol. 2, 2005, 60–5.

[23] J. Cai, Z. Cao, and L. Zhang, "Learning a single tucker decomposition network for lossy image compression with multiple bits-per-pixel rates", in *IEEE Transactions on Image Processing*, Vol. 29, 2020, 3612–25.

[24] J. Cao, X. Yao, H. Zhang, J. Jin, Y. Zhang, and B. W.-K. Ling, “Slimmable multi-task image compression for human and machine vision”, in *IEEE Access*, Vol. 11, 2023, 29946–58.

[25] H.-S. Chang, E. Learned-Miller, and A. McCallum, “Active bias: Training more accurate neural networks by emphasizing high variance samples”, in *Advances in Neural Information Processing Systems*, 2017, 1002–12.

[26] Y.-H. Chen, Y.-C. Weng, C.-H. Kao, C. Chien, W.-C. Chiu, and W.-H. Peng, “Transtic: Transferring transformer-based image compression from human perception to machine perception”, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, 23297–307.

[27] J. Chen, C. Yao, M. Liu, and Y. Zhao, “An End-to-End Mutual Enhancement Network Toward Image Compression and Semantic Segmentation”, in *Chinese Conference on Pattern Recognition and Computer Vision*, 2021, 623–35.

[28] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs”, in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 40, No. 4, 2017, 834–48.

[29] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, “Encoder-decoder with atrous separable convolution for semantic image segmentation”, in *Proceedings of the European Conference on Computer Vision*, 2018, 801–18.

[30] S. Chen, J. Jin, L. Meng, W. Lin, Z. Chen, T.-S. Chang, Z. Li, and H. Zhang, “A new image codec paradigm for human and machine uses”, in *arXiv preprint arXiv:2112.10071*, 2021.

[31] T. Chen, H. Liu, Z. Ma, Q. Shen, X. Cao, and Y. Wang, “End-to-End Learnt Image Compression via Non-Local Attention Optimization and Improved Context Modeling”, in *IEEE Transactions on Image Processing*, Vol. 30, 2021, 3179–91.

[32] Y. Chen, D. Murherjee, J. Han, A. Grange, Y. Xu, Z. Liu, S. Parker, C. Chen, H. Su, U. Joshi, *et al.*, “An overview of core coding tools in the AV1 video codec”, in *2018 picture coding symposium*, 2018, 41–5.

[33] Z. Chen, L.-Y. Duan, S. Wang, W. Lin, and A. C. Kot, “Data representation in hybrid coding framework for feature maps compression”, in *Proceedings of the IEEE International Conference on Image Processing*, 2020, 3094–8.

[34] Z. Chen, K. Fan, S. Wang, L.-Y. Duan, W. Lin, and A. Kot, “Lossy intermediate deep learning feature compression and evaluation”, in *ACM Trans. Multimedia*, 2019, 2414–22.

[35] Z. Cheng, H. Sun, M. Takeuchi, and J. Katto, "Deep Convolutional AutoEncoder-based Lossy Image Compression", in *Proceedings of 2018 Picture Coding Symposium*, 2018, 253–7.

[36] Z. Cheng, H. Sun, M. Takeuchi, and J. Katto, "Learned image compression with discretized gaussian mixture likelihoods and attention modules", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, 7939–48.

[37] H. Choi and I. V. Bajić, "Latent-space scalability for multi-task collaborative intelligence", in *2021 IEEE International Conference on Image Processing*, 2021, 3562–6.

[38] H. Choi and I. V. Bajić, "Scalable image coding for humans and machines", in *IEEE Transactions on Image Processing*, Vol. 31, 2022, 2739–54.

[39] H. Choi and I. V. Bajić, "Scalable video coding for humans and machines", in *2022 IEEE 24th International Workshop on Multimedia Signal Processing*, 2022, 1–6.

[40] Y. Choi, M. El-Khamy, and J. Lee, "Variable rate deep image compression with a conditional autoencoder", in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, 3146–54.

[41] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, 3213–23.

[42] M. Covell, N. Johnston, D. Minnen, S. J. Hwang, J. Shor, S. Singh, D. Vincent, and G. Toderici, "Target-Quality Image Compression with Recurrent, Convolutional Neural Networks", in *arXiv preprint arXiv:1705.06687*, 2017.

[43] Z. Cui, J. Wang, S. Gao, T. Guo, Y. Feng, and B. Bai, "Asymmetric Gained Deep Image Compression With Continuous Rate Adaptation", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, 10532–41.

[44] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks", in *Proceedings of the Advances in Neural Information Processing Systems*, 2016, 379–87.

[45] P. Dendorfer, H. Rezatofighi, A. Milan, J. Shi, D. Cremers, I. Reid, S. Roth, K. Schindler, and L. Leal-Taixé, "Mot20: A benchmark for multi object tracking in crowded scenes", in *arXiv preprint arXiv:2003.09003*, 2020.

[46] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database", in *2009 IEEE conference on computer vision and pattern recognition*, 2009, 248–55.

[47] Y. Ding, L. Wang, D. Fan, and B. Gong, “A semi-supervised two-stage approach to learning from noisy labels”, in *2018 IEEE Winter Conference on Applications of Computer Vision*, 2018, 1215–24.

[48] Y. Dong and W. D. Pan, “A survey on compression domain image and video data processing and analysis techniques”, in *Information*, Vol. 14, No. 3, 2023, 184.

[49] L.-Y. Duan, V. Chandrasekhar, J. Chen, J. Lin, Z. Wang, T. Huang, B. Girod, and W. Gao, “Overview of the MPEG-CDVS standard”, in *IEEE Transactions on Image Processing*, Vol. 25, No. 1, 2015, 179–94.

[50] L.-Y. Duan, Y. Lou, Y. Bai, T. Huang, W. Gao, V. Chandrasekhar, J. Lin, S. Wang, and A. C. Kot, “Compact descriptors for video analysis: The emerging MPEG standard”, in *IEEE MultiMedia*, Vol. 26, No. 2, 2018, 44–54.

[51] L. Duan, J. Liu, W. Yang, T. Huang, and W. Gao, “Video coding for machines: A paradigm of collaborative compression and intelligent analytics”, in *IEEE Transactions on Image Processing*, Vol. 29, 2020, 8680–95.

[52] T. Dumas, A. Roumy, and C. Guillemot, “Autoencoder Based Image Compression: Can the Learning be Quantization Independent?”, in *Proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018, 1188–92.

[53] T. Dumas, A. Roumy, and C. Guillemot, “Image compression with Stochastic Winner-Take-All Auto-Encoder”, in *Proceedings of 2017 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017, 1512–6.

[54] X. Fang, Y. Duan, Q. Du, X. Tao, and F. Li, “Sketch assisted face image coding for human and machine vision: a joint training approach”, in *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[55] Z. Fang, L. Shen, M. Li, Z. Wang, and Y. Jin, “Priors guided extreme underwater image compression for machine vision and human vision”, in *IEEE Journal of Oceanic Engineering*, 2023.

[56] R. Feng, Y. Gao, X. Jin, R. Feng, and Z. Chen, “Semantically structured image compression via irregular group-based decoupling”, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, 17237–47.

[57] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, “DSSD: Deconvolutional Single Shot Detector”, in *arXiv preprint arXiv:1701.06659*, 2017.

[58] L. Galteri, L. Seidenari, M. Bertini, and A. D. Bimbo, “Deep Generative Adversarial Compression Artifact Removal”, in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, 4826–35.

[59] C. Gao, D. Liu, L. Li, and F. Wu, "Towards task-generic image compression: A study of semantics-oriented metrics", in *IEEE Transactions on Multimedia*, Vol. 25, 2021, 721–35.

[60] K. Gregor, F. Besse, D. J. Rezende, I. Danihelka, and D. Wierstra, "Towards Conceptual Compression", in *Proceedings of Advances In Neural Information Processing Systems*, 2016, 3549–57.

[61] S. Guo, W. Huang, H. Zhang, C. Zhuang, D. Dong, M. R. Scott, and D. Huang, "Curriculumnet: Weakly supervised learning from large-scale web images", in *Proceedings of the European Conference on Computer Vision*, 2018, 135–50.

[62] S. Guo, Z. Chen, Y. Zhao, N. Zhang, X. Li, and L. Duan, "Toward Scalable Image Feature Compression: A Content-Adaptive and Diffusion-Based Approach", in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, 1431–42.

[63] Z. Guo, Z. Zhang, R. Feng, and Z. Chen, "Causal Contextual Prediction for Learned Image Compression", in *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[64] H. Hadizadeh and I. V. Bajić, "Learned Scalable Video Coding For Humans and Machines", in *arXiv preprint arXiv:2307.08978*, 2023.

[65] B. Han, I. W. Tsang, L. Chen, P. Y. Celina, and S.-F. Fung, "Progressive stochastic learning for noisy labels", in *IEEE transactions on neural networks and learning systems*, No. 99, 2018, 1–13.

[66] B. Han, Q. Yao, X. Yu, G. Niu, M. Xu, W. Hu, I. Tsang, and M. Sugiyama, "Co-teaching: Robust training of deep neural networks with extremely noisy labels", in *Advances in Neural Information Processing Systems*, 2018, 8527–37.

[67] J. Han, D. Zhang, G. Cheng, N. Liu, and D. Xu, "Advanced deep-learning techniques for salient and category-specific object detection: A survey", in *IEEE Signal Processing Magazine*, Vol. 35, No. 1, 2018, 84–100.

[68] Z. Hayder, X. He, and M. Salzmann, "Boundary-aware instance segmentation", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, 5696–704.

[69] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN", in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, 2961–9.

[70] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, 770–8.

[71] X. He, Q. Hu, X. Zhang, C. Zhang, W. Lin, and X. Han, "Enhancing HEVC Compressed Videos with a Partition-Masked Convolutional Neural Network", in *Proceedings of 2018 25th IEEE International Conference on Image Processing*, 2018, 216–20.

[72] T. M. Hoang, J. Zhou, and Y. Fan, “Image compression with encoder-decoder matched semantic segmentation”, in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020, 160–1.

[73] M. Hong, X. Zhang, G. Li, and Q. Huang, “Fine-grained feature generation for generalized zero-shot video classification”, in *IEEE Transactions on Image Processing*, Vol. 32, 2023, 1599–612.

[74] M. Hong, X. Zhang, G. Li, and Q. Huang, “Multi-modal multi-grained embedding learning for generalized zero-shot video classification”, in *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[75] A. Hore and D. Ziou, “Image quality metrics: PSNR vs. SSIM”, in *2010 20th international conference on pattern recognition*, 2010, 2366–9.

[76] Y. Hou, L. Zheng, and S. Gould, “Learning to structure an image with few colors”, in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, 2020, 10113–22.

[77] Y. Hu, W. Yang, M. Li, and J. Liu, “Progressive Spatial Recurrent Neural Network for Intra Prediction”, in *IEEE Transactions on Multimedia*, Vol. 21, No. 12, 2019, 3024–37.

[78] Y. Hu, W. Yang, and J. Liu, “Coarse-to-fine hyper-prior modeling for learned image compression”, in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34, No. 07, 2020, 11013–20.

[79] Y. Hu, W. Yang, Z. Ma, and J. Liu, “Learning end-to-end lossy image compression: A benchmark”, in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[80] Y. Hu, S. Yang, W. Yang, L.-Y. Duan, and J. Liu, “Towards coding for human and machine vision: A scalable image coding approach”, in *2020 IEEE International Conference on Multimedia and Expo*, 2020, 1–6.

[81] G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, “Densely connected convolutional networks”, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, 4700–8.

[82] H. Huang, W. Yang, W. Xiang, J. Liu, and L.-Y. Duan, “Collaborative scalable visual compression for human-centered videos”, in *2022 IEEE International Symposium on Circuits and Systems*, 2022, 2988–92.

[83] J. Huang *et al.*, “Speed/accuracy trade-offs for modern convolutional object detectors”, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, 7310–1.

[84] J. Huang, L. Qu, R. Jia, and B. Zhao, “O2u-net: A simple noisy label detection approach for deep neural networks”, in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, 3326–34.

[85] Z. Huang, C. Jia, S. Wang, and S. Ma, “Hmfvc: A human-machine friendly video compression scheme”, in *IEEE Transactions on Circuits and Systems for Video Technology*, 2022.

[86] A. Ikusan and R. Dai, “Deep Feature Compression with Rate-Distortion Optimization for Networked Camera Systems”, in *Proceedings of the 14th Conference on ACM Multimedia Systems*, 2023, 86–96.

[87] D. J. Im, C. D. Kim, H. Jiang, and R. Memisevic, “Generating images with recurrent adversarial networks”, in *CoRR, vol. abs/1602.05110*, 2016.

[88] K. Islam, L. M. Dang, S. Lee, and H. Moon, “Image Compression With Recurrent Neural Network and Generalized Divisive Normalization”, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, 1875–9.

[89] S. Jamil, M. J. Piran, M. Rahman, and O.-J. Kwon, “Learning-driven lossy image compression: A comprehensive survey”, in *Engineering Applications of Artificial Intelligence*, Vol. 123, 2023, 106361.

[90] H. Jégou, M. Douze, C. Schmid, and P. Pérez, “Aggregating local descriptors into a compact image representation”, in *Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition*, 2010, 3304–11.

[91] J. Jiang, J. Ma, Z. Wang, C. Chen, and X. Liu, “Hyperspectral image classification in the presence of noisy labels”, in *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 57, No. 2, 2019, 851–65.

[92] L. Jiang, Z. Zhou, T. Leung, L.-J. Li, and L. Fei-Fei, “Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels”, in *International Conference on Machine Learning*, 2018, 2304–13.

[93] X. Jin, R. Feng, S. Sun, R. Feng, T. He, and Z. Chen, “Semantical video coding: Instill static-dynamic clues into structured bitstream for ai tasks”, in *Journal of Visual Communication and Image Representation*, Vol. 93, 2023, 103816.

[94] N. Johnston, D. Vincent, D. Minnen, M. Covell, S. Singh, T. Chinen, S. J. Hwang, J. Shor, and G. Toderici, “Improved Lossy Image Compression with Priming and Spatially Adaptive Bit Rates for Recurrent Networks”, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, 4385–93.

[95] B. Kang, S. Tripathi, and T. Q. Nguyen, “Toward Joint Image Generation and Compression using Generative Adversarial Networks”, in *arXiv preprint arXiv:1901. 07838*, 2019.

[96] A. C. Karaca and M. K. Güllü, “Target preserving hyperspectral image compression using weighted PCA and JPEG2000”, in *Proceedings of International Conference on Image and Signal Processing*, 2018, 508–16.

[97] W. Khalaf, D. Zaghar, and N. Hashim, “Enhancement of Curve-Fitting Image Compression Using Hyperbolic Function”, in *Symmetry*, Vol. 11, 2019, 291.

[98] E. Kodak, "Kodak Lossless True Color Image Suite (PhotoCD PCD0992)", 1992.

[99] F. Kong, K. Hu, Y. Li, D. Li, and D. Zhao, "Spectral-Spatial Feature Partitioned Extraction Based on CNN for Multispectral Image Compression", in *Remote Sensing*, Vol. 13, No. 1, 2020, 9.

[100] T. Kong, F. Sun, A. Yao, H. Liu, M. Lu, and Y. Chen, "RON: Reverse connection with objectness prior networks for object detection", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, 5936–44.

[101] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther, "Autoencoding beyond pixels using a learned similarity metric", in *Proceedings of The 33rd International Conference on Machine Learning*, Vol. 48, 2016, 1558–66.

[102] J. Lee, S. Cho, and M. Kim, "An end-to-end joint learning scheme of image compression and quality enhancement with improved entropy minimization", in *arXiv preprint arXiv:1912.12817*, 2020.

[103] Y. Lee and J. Park, "Centermask: Real-time anchor-free instance segmentation", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, 13906–15.

[104] Z. Lei, P. Duan, X. Hong, J. F. Mota, J. Shi, and C.-X. Wang, "Progressive deep Image compression for hybrid contexts of image classification and reconstruction", in *IEEE Journal on Selected Areas in Communications*, Vol. 41, No. 1, 2022, 72–89.

[105] Z. Lei, X. Hong, J. Shi, M. Su, C. Lin, and W. Xia, "Quantization-Based Adaptive Deep Image Compression Using Semantic Information", in *IEEE Access*, 2023.

[106] "SSD: Single shot multibox detector", in *Computer Vision–ECCV*, ed. B. Leibe, J. Matas, N. Sebe, and M. Welling, 2016, 21–37.

[107] J. Li and Z. Liu, "Multispectral transforms using convolution neural networks for remote sensing multispectral image compression", in *Remote Sensing*, Vol. 11, No. 7, 2019, 759.

[108] M. Li, K. Zhang, J. Li, W. Zuo, R. Timofte, and D. Zhang, "Learning Context-Based Nonlocal Entropy Modeling for Image Compression", in *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[109] W. Li, W. Sun, Y. Zhao, Z. Yuan, and Y. Liu, "Deep Image Compression with Residual Learning", in *Applied Sciences*, Vol. 10, No. 11, 2020, 4023.

[110] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng, "HDenseUNet: hybrid densely connected UNet for liver and tumor segmentation from CT volumes", in *IEEE Transactions on Medical Imaging*, Vol. 37, No. 12, 2018, 2663–74.

[111] Y. Li, C. Jia, S. Wang, X. Zhang, S. Wang, S. Ma, and W. Gao, “Joint rate-distortion optimization for simultaneous texture and deep feature compression of facial images”, in *2018 IEEE fourth international conference on multimedia big data (BigMM)*, 2018, 1–5.

[112] Z. Li, C. Peng, G. Yu, X. Zhang, Y. Deng, and J. Sun, “DetNet: A backbone network for object detection”, in *arXiv preprint arXiv:1804.06215*, 2018.

[113] G. Lin, A. Milan, C. Shen, and I. Reid, “RefineNet: Multi-path refinement networks for high-resolution semantic segmentation”, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, 1925–34.

[114] H. Lin, B. Chen, Z. Zhang, J. Lin, X. Wang, and T. Zhao, “DeepSVC: Deep scalable video coding for both machine and human vision”, in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, 9205–14.

[115] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature pyramid networks for object detection”, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, 2117–25.

[116] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, “Focal loss for dense object detection”, in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, 2980–8.

[117] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft coco: Common objects in context”, in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 2014, 740–55.

[118] D. Liu, B. Wen, Y. Fan, C. C. Loy, and T. S. Huang, “Non-local recurrent network for image restoration”, in *Neural Information Processing Systems*, 2018, 1673–82.

[119] H. Liu, T. Chen, Q. Shen, T. Yue, and Z. Ma, “Deep Image Compression via End-to-End Learning”, in *Proceedings of Computer Vision Pattern Recognition*, Vol. 06, 2018.

[120] J. Liu, H. Sun, and J. Katto, “Improving multiple machine vision tasks in the compressed domain”, in *2022 26th International Conference on Pattern Recognition*, 2022, 331–7.

[121] K. Liu, D. Liu, L. Li, N. Yan, and H. Li, “Semantics-to-signal scalable image compression with learned revertible representations”, in *International Journal of Computer Vision*, Vol. 129, No. 9, 2021, 2605–21.

[122] L. Liu, Z. Hu, Z. Chen, and D. Xu, “Icmh-net: Neural image compression towards both machine vision and human vision”, in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, 8047–56.

[123] L. Liu, T. Chen, H. Liu, S. Pu, L. Wang, and Q. Shen, “2C-Net: integrate image compression and classification via deep neural network”, in *Multimedia Systems*, Vol. 29, No. 3, 2023, 945–59.

[124] S. Liu *et al.*, “Receptive field block net for accurate and fast object detection”, in *Proceedings of the European Conference on Computer Vision*, 2018, 385–400.

[125] D. G. Lowe, “Distinctive image features from scale-invariant keypoints”, in *Int. J. Comput. Vis.* Vol. 60, No. 2, 2004, 91–110.

[126] J. Luengo, S.-O. Shim, S. Alshomrani, A. Altalhi, and F. Herrera, “Cnc-nos: Class noise cleaning by ensemble filtering and noise scoring”, in *Knowledge-Based Systems*, Vol. 140, 2018, 27–49.

[127] Y. Lyu and I. W. Tsang, “Curriculum Loss: Robust Learning and Generalization against Label Corruption”, in *arXiv preprint arXiv:1905.10045*, 2019.

[128] S. Ma, X. Zhang, S. Wang, X. Zhang, C. Jia, and S. Wang, “Joint feature and texture coding: Toward smart video representation via front-end intelligence”, in *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 29, No. 10, 2018, 3095–105.

[129] S. Ma, X. Zhang, C. Jia, Z. Zhao, S. Wang, and S. Wang, “Image and video compression with neural networks: A review”, in *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 30, No. 6, 2019, 1683–98.

[130] E. Malach and S. Shalev-Shwartz, “Decoupling” when to update” from” how to update””, in *Advances in Neural Information Processing Systems*, 2017, 960–70.

[131] Q. Mao, C. Wang, M. Wang, S. Wang, R. Chen, L. Jin, and S. Ma, “Scalable Face Image Coding via StyleGAN Prior: Towards Compression for Human-Machine Collaborative Vision”, in *IEEE Transactions on Image Processing*, 2023.

[132] Y. Mao, P. Chen, S. Wang, S. Wang, and D. Wu, “Peering into the sketch: Ultra-low bitrate face compression for joint human and machine perception”, in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, 2564–72.

[133] M. W. Marcellin, M. J. Gormish, A. Bilgin, and M. P. Boliek, “An overview of JPEG-2000”, in *Proceedings DCC 2000. Data compression conference*, 2000, 523–41.

[134] F. D. Martino, I. Perfilieva, and S. Sessa, “A Fast Multilevel Fuzzy Transform Image Compression Method”, in *Axioms*, Vol. 8, No. 4, 2019, 135.

[135] F. D. Martino and S. Sessa, “Multi-level fuzzy transforms image compression”, in *Journal of Ambient Intelligence and Humanized Computing*, Vol. 10, No. 7, 2019, 2745–56.

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

[137] D. Minnen, J. Balle, and G. D. Toderici, "Joint autoregressive and hierarchical priors for learned image compression", in *Proceedings of Advances in Neural Information Processing Systems*, 2018, 10771–80.

[138] D. Minnen and S. Singh, "Channel-Wise Autoregressive Entropy Models for Learned Image Compression", in *Proceedings of 2020 IEEE International Conference on Image Processing*, 2020, 3339–43.

[139] D. T. Nguyen, C. K. Mummadi, T. P. N. Ngo, T. H. P. Nguyen, L. Beggel, and T. Brox, "SELF: Learning to Filter Noisy Labels with Self-Ensembling", in *arXiv preprint arXiv:1910.01842*, 2019.

[140] D. T. Nguyen, T.-P.-N. Ngo, Z. Lou, M. Klar, L. Beggel, and T. Brox, "Robust Learning Under Label Noise With Iterative Noise-Filtering", in *arXiv preprint arXiv:1906.00216*, 2019.

[141] C. G. Northcutt, T. Wu, and I. L. Chuang, "Learning with confident examples: Rank pruning for robust classification with noisy labels", in *Uncertainty in Artificial Intelligence - Proceedings of the 33rd Conference, UAI 2017*, 2017.

[142] J.-R. Ohm, G. J. Sullivan, H. Schwarz, T. K. Tan, and T. Wiegand, "Comparison of the coding efficiency of video coding standards—including high efficiency video coding (HEVC)", in *IEEE Transactions on circuits and systems for video technology*, Vol. 22, No. 12, 2012, 1669–84.

[143] V. A. de Oliveira, M. Chabert, T. Oberlin, C. Poulliat, M. Bruno, C. Latry, M. Carlavan, S. Henrot, F. Falzon, and R. Camarero, "Reduced-Complexity End-to-End Variational Autoencoder for on Board Satellite Image Compression", in *Remote Sensing*, Vol. 13, No. 3, 2021, 447.

[144] Y. Ollivier, "Auto-encoders: reconstruction versus compression", 2014.

[145] A. G. Ororbia, A. Mali, J. Wu, S. O'Connell, W. Dreese, D. Miller, and C. L. Giles, "Learned Neural Iterative Decoding for Lossy Image Compression Systems", in *Proceedings of 2019 Data Compression Conference*, 2019, 3–12.

[146] W. Ouyang, K. Wang, X. Zhu, and X. Wang, "Chained cascade network for object detection", in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, 1956–64.

[147] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, "Enet: A deep neural network architecture for real-time semantic segmentation", in *arXiv preprint arXiv:1606. 02147*, 2016.

[148] F. Perronnin, Y. Liu, J. Sánchez, and H. Poirier, "Large-scale image retrieval with compressed fisher vectors", in *Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition*, 2010, 3384–91.

[149] P. O. Pinheiro, R. Collobert, and P. Dollár, “Learning to segment object candidates”, in *Advances in Neural Information Processing Systems*, 2015, 1990–8.

[150] A. Prakash, N. Moran, S. Garber, A. DiLillo, and J. Storer, “Semantic Perceptual Image Compression using Deep Convolution Networks”, in *Proceedings of 2017 Data Compression Conference*, 2017, 250–9.

[151] A. Punnappurath and M. S. Brown, “Learning raw image reconstruction-aware deep image compressors”, in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 42, No. 4, 2020, 1013–9.

[152] P. Purkait, C. Zhao, and C. Zach, “SPP-net: Deep absolute pose regression with synthetic views”, in *arXiv preprint arXiv:1712.03452*, 2017.

[153] S. K. Raman, A. Ramesh, V. Naganoor, S. Dash, G. Kumaravelu, and H. Lee, “CompressNet: Generative Compression at Extremely Low Bitrates”, in *Proceedings of The IEEE Winter Conference on Applications of Computer Vision*, 2020, 2325–33.

[154] S. Reed, H. Lee, D. Anguelov, C. Szegedy, D. Erhan, and A. Rabinovich, “Training deep neural networks on noisy labels with bootstrapping”, in *arXiv preprint arXiv:1412.6596*, 2014.

[155] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, “Generalized intersection over union: A metric and a loss for bounding box regression”, in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, 658–66.

[156] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation”, in *International Conference on Medical Image Computing and Computer-assisted Intervention*, 2015, 234–41.

[157] A. Sento, “Image Compression with Auto-encoder Algorithm using Deep Neural Network (DNN)”, in *Proceedings of 2016 Management and Innovation Technology International Conference*, 2016, 99.

[158] R. R. Shamir, Y. Duchin, J. Kim, G. Sapiro, and N. Harel, “Continuous dice coefficient: a method for evaluating probabilistic segmentations”, in *arXiv preprint arXiv:1906.11031*, 2019.

[159] Z. Shen, Z. Liu, J. Li, Y.-G. Jiang, Y. Chen, and X. Xue, “DSOD: Learning deeply supervised object detectors from scratch”, in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, 1919–27.

[160] X. Sheng, L. Li, D. Liu, and H. Li, “VNVC: A Versatile Neural Video Coding Framework for Efficient Human-Machine Vision”, in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[161] S. Singh, S. Abu-El-Haija, N. Johnston, J. Ballé, A. Shrivastava, and G. Toderici, “End-to-end learning of compressible features”, in *Proceedings of the IEEE International Conference on Image Processing*, 2020, 3349–53.

[162] J. Sivic and A. Zisserman, “Video Google: A text retrieval approach to object matching in videos”, in *Proceedings of the IEEE International Conference on Computer Vision*, 2003, 1470–7.

[163] J. Song, T. He, L. Gao, X. Xu, A. Hanjalic, and T. H. Shen, “Unified Binary Generative Adversarial Network for Image Retrieval and Compression”, in *International Journal of Computer Vision*, Vol. 128, No. 8, 2020, 2243–64.

[164] K. Soomro, A. R. Zamir, and M. Shah, “UCF101: A dataset of 101 human actions classes from videos in the wild”, in *arXiv preprint arXiv:1212.0402*, 2012.

[165] S. Sun, T. He, and Z. Chen, “Semantic structured image coding framework for multiple intelligent applications”, in *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 31, No. 9, 2020, 3631–42.

[166] S. Suzuki, M. Takagi, K. Hayase, T. Onishi, and A. Shimizu, “Image pretransformation for recognition-aware image compression”, in *Proceedings of the IEEE International Conference on Image Processing*, 2019, 2686–90.

[167] S. Suzuki, M. Takagi, S. Takeda, R. Tanida, and H. Kimata, “Deep feature compression with spatio-temporal arranging for collaborative intelligence”, in *Proceedings of the IEEE International Conference on Image Processing*, 2020, 3099–103.

[168] D. Tellez, G. Litjens, J. van der Laak, and F. Ciompi, “Neural Image Compression for Gigapixel Histopathology Image Analysis”, in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 43, No. 2, 2021, 567–78.

[169] L. Theis, W. Shi, A. Cunningham, and F. Huszar, “Lossy image compression with compressive autoencoders”, in *Proceedings of International Conference on Learning Representations*, 2017.

[170] Y. Tian, G. Lu, Y. Yan, G. Zhai, L. Chen, and Z. Gao, “A Coding Framework and Benchmark towards Low-Bitrate Video Understanding”, in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[171] Y. Tian, G. Lu, G. Zhai, and Z. Gao, “Non-Semantics Suppressed Mask Learning for Unsupervised Video Semantic Compression”, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, 13610–22.

[172] G. Toderici, S. M. O'Malley, S. J. Hwang, D. Vincent, D. Minnen, S. Baluja, M. Covell, and R. Sukthankar, "Variable Rate Image Compression with Recurrent Neural Networks", in *CoRR, vol. abs/1511.06085*, 2016.

[173] G. Toderici, D. Vincent, N. Johnston, S. J. Hwang, D. Minnen, J. Shor, and M. Covell, "Full Resolution Image Compression with Recurrent Neural Networks", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, 5306–14.

[174] G. Toderici, L. Theis, N. Johnston, E. Agustsson, F. Mentzer, J. Ballé, W. Shi, and R. Timofte, "CLIC 2020: Challenge on Learned Image Compression", Sections 2, 6, 2020.

[175] R. Torfason, F. Mentzer, E. Augustsson, M. Tschannen, R. Timofte, and L. V. Gool, "Towards Image Understanding from Deep Compression Without Decoding", in *Proceedings of International Conference on Learning Representations*, 2018.

[176] R. Torfason, F. Mentzer, E. Agustsson, M. Tschannen, R. Timofte, and L. Van Gool, "Towards image understanding from deep compression without decoding", in *arXiv preprint arXiv:1803.06131*, 2018.

[177] H. Tu, L. Li, W. Zhou, and H. Li, "Semantic scalable image compression with cross-layer priors", in *Proceedings of the 29th ACM International conference on multimedia*, 2021, 4044–52.

[178] S. Vicente, J. Carreira, L. Agapito, and J. Batista, "Reconstructing pascal voc", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, 41–8.

[179] F. Visin, M. Ciccone, A. Romero, K. Kastner, K. Cho, Y. Bengio, M. Matteucci, and A. Courville, "Reseg: A recurrent neural network-based model for semantic segmentation", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016, 41–8.

[180] G. K. Wallace, "The JPEG still picture compression standard", in *Communications of the ACM*, Vol. 34, No. 4, 1991, 30–44.

[181] C. Wang, Y. Han, and W. Wang, "An End-to-End Deep Learning Image Compression Framework Based on Semantic Analysis", in *Applied Sciences*, Vol. 9, No. 17, 2019, 3580.

[182] P. Wang, P. Chen, Y. Yuan, D. Liu, Z. Huang, X. Hou, and G. Cottrell, "Understanding convolution for semantic segmentation", in *Winter Conference on Applications of Computer Vision*, 2018, 1451–60.

[183] R. J. Wang, X. Li, and C. X. Ling, "Pelee: A real-time object detection system on mobile devices", in *Proceedings of the Advances in Neural Information Processing Systems*, 2018, 1963–72.

[184] S. Wang, S. Wang, W. Yang, X. Zhang, S. Wang, S. Ma, and W. Gao, "Towards analysis-friendly face representation with scalable feature and texture compression", in *IEEE Transactions on Multimedia*, Vol. 24, 2021, 3169–81.

[185] S. Wang, S. Wang, X. Zhang, S. Wang, S. Ma, and W. Gao, "Scalable facial image compression with deep feature reconstruction", in *2019 IEEE International Conference on Image Processing*, 2019, 2691–5.

[186] S. Wang, Z. Wang, S. Wang, and Y. Ye, "End-to-end compression towards machine vision: Network architecture design and optimization", in *IEEE Open Journal of Circuits and Systems*, Vol. 2, 2021, 675–85.

[187] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks", in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2018, 7794–803.

[188] Z. Wang and A. C. Bovik, "A universal image quality index", in *IEEE signal processing letters*, Vol. 9, No. 3, 2002, 81–4.

[189] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment", in *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, Vol. 2, 2003, 1398–402.

[190] Z. Wang, F. Li, J. Xu, and P. C. Cosman, "Human–machine interaction-oriented image coding for resource-constrained visual monitoring in IoT", in *IEEE Internet of Things Journal*, Vol. 9, No. 17, 2022, 16181–95.

[191] T. Wiegand, G. J. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the H. 264/AVC video coding standard", in *IEEE Transactions on circuits and systems for video technology*, Vol. 13, No. 7, 2003, 560–76.

[192] L. Wu, K. Huang, and H. Shen, "A GAN-based tunable image compression system", in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, 2334–42.

[193] X. Wu, R. He, Z. Sun, and T. Tan, "A light CNN for deep face representation with noisy labels", in *IEEE Transactions on Information Forensics and Security*, Vol. 13, No. 11, 2018, 2884–96.

[194] Y. Wu, X. Li, Z. Zhang, X. Jin, and Z. Chen, "Learned Block-based Hybrid Image Compression", in *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[195] Y. Wu, P. An, C. Yang, and X. Huang, "Scalable image coding with enhancement features for human and machine", in *Multimedia Systems*, Vol. 30, No. 2, 2024, 77.

[196] Z. Wu, H. Wang, H. Wang, and Y. Zhang, "End to End Scalable Image Coding for Machines", in *2023 3rd International Conference on Intelligent Communications and Computing*, 2023, 217–21.

[197] S. Xia, K. Liang, W. Yang, L.-Y. Duan, and J. Liu, “An emerging coding paradigm VCM: A scalable coding approach beyond feature and signal”, in *2020 IEEE International Conference on Multimedia and Expo*, 2020, 1–6.

[198] Y. Xie, K. L. Cheng, and Q. Chen, “Enhanced invertible encoding for learned image compression”, in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, 162–70.

[199] Y. Xue and J. Su, “Attention Based Image Compression Post-Processing Convolutional Neural Network”, in *Proceedings of CVPR Workshops*, 2019.

[200] R. Yacouby and D. Axman, “Probabilistic extension of precision, recall, and f1 score for more thorough evaluation of classification models”, in *Proceedings of the first workshop on evaluation and comparison of NLP systems*, 2020, 79–91.

[201] R. J. Yadav and M. S. Nagmode, “Compression of hyperspectral image using PCA–DCT technology”, in *Proceedings of Innovations in Electronics and Communication Engineering*, 2018, 269–77.

[202] N. Yan, C. Gao, D. Liu, H. Li, L. Li, and F. Wu, “SSSIC: semantics-to-signal scalable image coding with learned structural representations”, in *IEEE Transactions on Image Processing*, Vol. 30, 2021, 8939–54.

[203] Y. Yan, Z. Xu, I. W. Tsang, G. Long, and Y. Yang, “Robust semi-supervised learning through label aggregation”, in *30th AAAI Conference on Artificial Intelligence, AAAI 2016*, 2016, 2244–50.

[204] M. Yang, K. Yu, C. Zhang, Z. Li, and K. Yang, “DenseASPP for semantic segmentation in street scenes”, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, 3684–92.

[205] W. Yang, H. Huang, J. Liu, and A. C. Kot, “Facial image compression via neural image manifold compression”, in *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[206] Z. Yang, Y. Wang, C. Xu, P. Du, C. Xu, C. Xu, and Q. Tian, “Discernible image compression”, in *ACM Trans. Multimedia*, 2020, 1561–9.

[207] X. Yi, H. Wang, S. Kwong, and C.-C. J. Kuo, “Task-driven video compression for humans and machines: Framework design and optimization”, in *IEEE Transactions on Multimedia*, 2022.

[208] T. M. Zeegers, D. M. Pelt, T. van Leeuwen, R. van Liere, and K. J. Batenburg, “Task-Driven Learned Hyperspectral Data Reduction Using End-to-End Supervised Deep Learning”, in *Journal of Imaging*, Vol. 6, No. 12, 2020, 132.

[209] J. Zhang, C. Jia, M. Lei, S. Wang, S. Ma, and W. Gao, “Recent development of AVS video coding standard: AVS3”, in *2019 picture coding symposium*, 2019, 1–5.

[210] X. Zhang and X. Wu, "Near-lossless L-infinity constrained Multi-rate Image Decompression via Deep Neural Network", 2018.

[211] X. Zhang, S. Ma, S. Wang, X. Zhang, H. Sun, and W. Gao, "A joint compression scheme of video feature descriptors and visual content", in *IEEE Transactions on Image Processing*, Vol. 26, No. 2, 2016, 633–47.

[212] Y. Zhang, K. Li, K. Li, B. Zhong, and Y. Fu, "Residual non-local attention networks for image restoration", in *International Conference on Learning Representations*, 2019.

[213] Y. Zhang, C. Jia, J. Chang, and S. Ma, "Machine perception-driven image compression: A layered generative approach", in *arXiv preprint arXiv:2304.06896*, 2023.

[214] Y. Zhang, L. Zhu, G. Jiang, S. Kwong, and C.-C. J. Kuo, "A survey on perceptually optimized video coding", in *ACM Computing Surveys*, Vol. 55, No. 12, 2023, 1–37.

[215] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, 2881–90.

[216] Z.-Q. Zhao, P. Zheng, S.-T. Xu, and X. Wu, "Object detection with deep learning: A review", in *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 30, No. 11, 2019, 3212–32.

[217] T. Zheng, W. Deng, and J. Hu, "Cross-age lfw: A database for studying cross-age face recognition in unconstrained environments", in *arXiv preprint arXiv:1708.08197*, 2017.

[218] L. Zhou, C. Cai, Y. Gao, S. Su, and J. Wu, "Variational Autoencoder for Low Bit-rate Image Compression", in *Proceedings of The IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018.