Task-Aware Encoder Control for Deep Video Compression

Xingtong Ge[1,2]*   Jixiang Luo[2]   Xinjie Zhang[3]   Tongda Xu[4]   Guo Lu[5]
Dailan He[6]   Jing Geng[1]‡   Yan Wang[4]   Jun Zhang[3]   Hongwei Qin[2]
[1]Beijing Institute of Technology   [2]SenseTime Research
[3]The Hong Kong University of Science and Technology
[4]Institute for AI Industry Research (AIR), Tsinghua University
[5]Shanghai Jiaotong University   [6]The Chinese University of Hong Kong

Abstract

Prior research on deep video compression (DVC) for machine tasks typically necessitates training a unique codec for each specific task, mandating a dedicated decoder per task. In contrast, traditional video codecs employ a flexible encoder controller, enabling the adaptation of a single codec to different tasks through mechanisms like mode prediction. Drawing inspiration from this, we introduce an innovative encoder controller for deep video compression for machines. This controller features a mode prediction and a Group of Pictures (GoP) selection module. Our approach centralizes control at the encoding stage, allowing for adaptable encoder adjustments across different tasks, such as detection and tracking, while maintaining compatibility with a standard pre-trained DVC decoder. Empirical evidence demonstrates that our method is applicable across multiple tasks with various existing pre-trained DVCs. Moreover, extensive experiments demonstrate that our method outperforms previous DVC by about 25% bitrate for different tasks, with only one pre-trained decoder.

1. Introduction

Over the past decades, video analysis techniques have proliferated across a variety of fields, including smart cities, autonomous vehicles, and traffic surveillance. For these applications, videos are often compressed before being transmitted to cloud-based systems for further machine vision analyses, such as object detection and tracking. However, current video compression techniques, which range from traditional codecs [2, 33, 42] to recent learning-based codecs [14, 18, 24, 25, 29], primarily cater to the human vi-
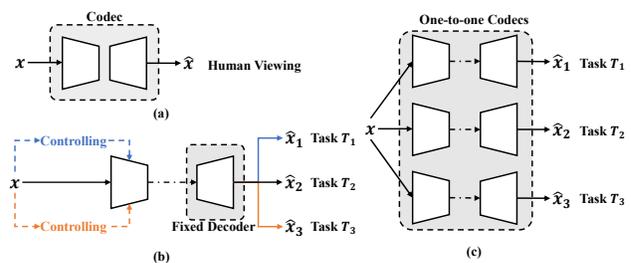
Figure 1. (a) Mainstream video codec that serves the human viewing. (b) Our controlled video codecs for machine vision with fixed decoder. (c) Other video codec for machine vison with one-to-one encoders and decoders.

sual system, as shown in Fig. 1(a). This specificity leads to inefficiencies, as machine vision tasks typically focus on selective semantic details and regions within the video frames, rather than the whole frames.

To tackle this problem, recent studies [1, 4, 11–13, 20, 38, 40, 44, 47] have explored scalable compression for multiple tasks through different layers in image and video coding. They propose a base layer dedicated to machine vision, with an enhancement layer containing additional information for human vision. Additionally, the latest work [36] has tried to compress semantic features at the encoder side, which serve as side information to supplement decoded images with more semantic details for machine vision at lower bitrates. Despite these advances, these methods require individually customized codecs when applied to different downstream tasks, as the "One-to-one Codecs" shown in Fig. 1(c). In contrast, traditional codecs can control the encoding process for different objectives, like PSNR and SSIM optimization, using a single decoder. Inspired by this feature, **in this paper, we focus on how to use one pre-trained DVC-decoder to support both human and multiple machine vision tasks**.

In both residual [14, 15, 24] and conditional [18, 19, 29] DVC series, the encoded bitstreams predominantly comprise two parts: motion and residual/contextual informa-

tion. It is observed that a substantial portion of the bitrate, often exceeding 80% in high bitrate models, is allocated to the residual/contextual information. This allocation is primarily for high-quality frame reconstruction. However, this approach is inappropriate for downstream vision tasks such as tracking and detection, which primarily concentrate on the objects and their movements rather than the full frames.

To mitigate this issue, we introduce a dynamic vision mode prediction (DVMP) module specifically tailored for machine vision tasks. This module optimizes the entropy coding process by dynamically predicting the skip/no-skip coding mode for each feature element, based on its relevance to machine vision. Specifically, it takes hyperprior data from motion or residual/contextual features within the DVC framework as the input and assesses the utility of each feature element for machine vision. By replacing non-essential elements with their predicted mean value derived from the hyperprior network, our method effectively reduces bit consumption and circumvent the entropy decoding step for these elements to expedite the decoding process. In this case, we generate a novel frame type that significantly reduces bitrate usage compared to traditional P frames in the Group of Pictures (GoP) structure, while maintaining the information valuable for machine vision tasks. This new frame is designated as the P_m frame.

At the same time, it is important for DVC to maintain reconstruction quality because there exists significant reference relationship between frames, which is a crucial difference between image and video compression. Since our P_m frame degrades the fidelity of reconstruction compared to the P frame, we explore the reorganization of the Group of Pictures (GoP) structure. To be specific, we firstly explore to use a hand-crafted GoP structure, which already achieves significant rate-precision improvements. In the revised structure, both P frames and P_m frames derive their prediction from the preceding I/P frame. Further, we introduce a GoP selection network, which can dynamically determine the GoP structure during the encoding process and achieves a better rate-precision trade-off.

We choose three representative residual and conditional DVC methods as touchstones in our proposed video coding for machine vision framework. It's worth mentioning that our method utilizes pre-trained DVCs without altering the weights of the original encoders and decoders. Our control happens at encoding stage, leaving the decoders' architecture intact. Moreover, when human viewing is required, we can switch back to the original encoding procedure to restore reconstruction performance. In essence, using the proposed method, we can control the encoder of a DVC to adapt for both machine and human vision requirements.

The main contributions of our work are summarised as follows: (1) Built upon the mainstream DVC codecs, we propose a novel video coding for machine vision framework

that controls the encoder to adapt for different downstream vision tasks, such as video object detection and multi-object tracking. (2) We employ a dynamic vision mode prediction approach to refine the original P frame, effectively reducing the bitrate while preserving critical information pertinent to vision tasks. Furthermore, we utilize a GoP selection strategy to dynamically forecast the coding GoP structure during the encoding stage, which controls the bitstream for better rate-precision trade-off in downstream vision tasks. (3) Experiments show that our controlling method is novel and flexible for different DVC codecs, achieving up to more than 25% bitrate savings in various downstream tasks.

2. Related Works

2.1. Video Compression

Previous video codecs are designed to remove spatial-temporal redundancies effectively. Traditional video coding standards, including H.264/AVC [42], H.265/HEVC [33], and H.266/VVC [2], significantly increase the compression efficiency of images and videos. Recently, a number of learning-based video codecs [14, 18, 24, 25, 29, 35] have been proposed, designed based on residual coding or conditional coding, achieved better and better pixel-wise signal quality metrics, e.g., PSNR and MS-SSIM [41], which mainly serve the human visual experience. Recently, there are also some generative video coding methods [27, 45] that mainly consider visual comfort and perceptual quality.

2.2. Compression for Machine Vision

To facilitate the efficiency of machine vision tasks, early research works were devoted to extracting visual features from signals. Early standards, such as CDVA [8] and CDVS [7], suggest the pre-extraction and transportation of image keypoints to facilitate image indexing or retrieval tasks. With the development of deep learning, some studies began to explore connections between compression and downstream tasks, some studies [1, 3, 4, 11, 13, 20, 38, 40, 44, 47] also focus on the joint optimization of image compression and downstream machine vision tasks by introducing a rate-distortion optimization strategy guided by downstream tasks or by adding a task-specific feature encoding stream. For instance, Torfason et al. [38] introduced a method for executing image understanding tasks, such as classification and segmentation, on compressed outputs from learning-based image compression techniques. Furthermore, Lu et al. [26] enhanced traditional codecs with a preprocessing step, thereby improving codec performance for downstream vision tasks. The field has also witnessed the emergence of self-supervised representation learning methods aimed at deriving compact semantic representations. Dubois et al. [9] presented a theoretical framework suggesting that the distortion term in the lossy rate-
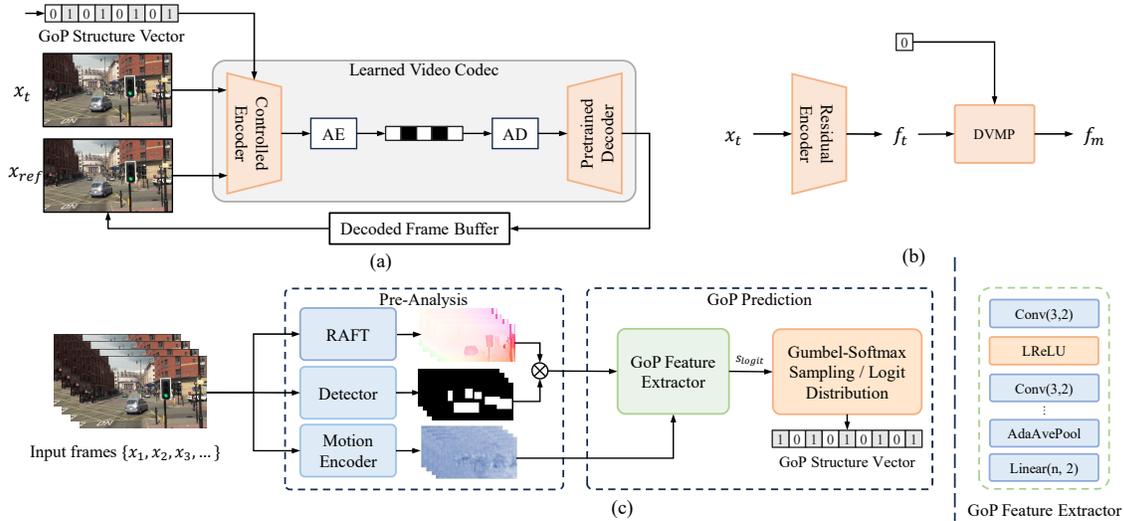
Figure 2. (a) Overview of our "Controlling DVC for Machine" framwork. Given an input GoP, we firstly use GoP Selection network to predict the GoP sructure, then the predicted structure controls the encoding procedure to encode input frames for machine vision tasks. (b) The "0" element controls encoder to use DVMP. (c) The GoP Selection network, including the pre-analysis stage and GoP prediction stage.

distortion trade-off for image classification could be approximated by a contrastive learning objective. Feng et al. [10] proposed a method to learn a unified feature representation for AI tasks from unlabeled data. These methodologies necessitate fine-tuning the downstream models to adapt to the learned features.

Despite advancements in image coding for machine applications, reconstructing high-fidelity video from extracted features remains a formidable challenge due to the critical inter-frame reference relationships. Addressing the requirements of both machine and human vision, Tian et al. [36] introduced a self-supervised edge representation as a semantic intermediary, which preserves the video's semantic structure. Furthermore, Lin et al. [21] developed a scalable video coding framework tailored for machines, segregating semantic features for machine analysis and human viewing into separate bitstreams.

Most existing approaches [21, 36, 37]necessitate the use of distinct encoders and decoders for various tasks, which complicates and burdens the coding pipelines. Our methodology advances the "Coding for Machine" paradigm by focusing on the encoder side. We present a versatile framework capable of accommodating both human and machine vision demands without the need to modify the pre-trained decoders of existing video codecs.

2.3. Compressed Video Analysis/Understanding

There are also amounts of works performing video analysis tasks [17, 23, 39], such as image recognition, action recognition [31, 43] and multiple object tracking (MOT) [6, 16], in the compressed video domain. However, these methods focus on developing video analysis models that better lever-

age the partially decoded video stream, such as the motion vector. In contrast, our work focuses on the coding procedure, specifically the encoding procedure.

3. Method

3.1. Overview

Our coding framework is illustrated in Fig. 2 (a). Let $\mathcal{X} = \{x_1, \cdots, x_T\}$ represent the video sequence. The framework categorizes the P frames into two types: original P frames and machine vision-specific P_m frames. The categorization of frame x_t is determined by the GoP Selection Network, where a value of 1 signifies a P frame and 0 a P_m frame. The P_m frames, designed to reduce the bitrate while preserving critical information for future vision tasks, are generated from the preceding P frames using Dynamic Vision Mode Prediction (DVMP). To maintain the quality of the decoded video sequence, each P frame is computed from the previous P frame with superior visual quality, rather than from a P_m frame. In scenarios requiring coding for downstream vision tasks, a machine-centric control applies the GoP selection and DVMP, altering the encoding structure to "$I, P_m, P, \cdots$". Conversely, for human viewing, a human-centric control restores the encoding structure to its original form, "$I, P, P, P, \cdots$".

3.2. Dynamic Vision Mode Prediction

In our coding framework, we aim to minimize the video sequence bitrate. Initial analyses of residual video codecs reveal that they allocate a substantial portion of the bitrate—exceeding 80% in high bitrate configurations—to encode and transmit residual information. This approach en-
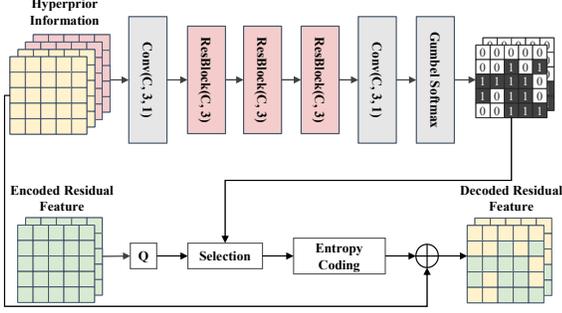
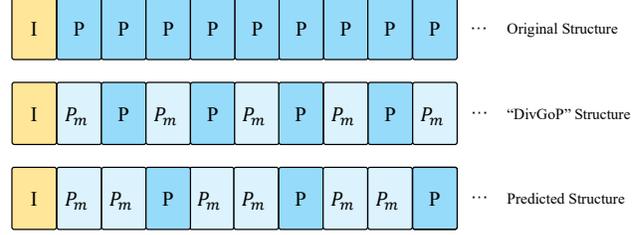Figure 3. Hyper-prior guided Dynamic Vision Mode Prediction network.

Figure 4. Different coding GoP structures. The original structure consists of I and P frames. In the middle, the hand-crafted structure for machine vision consists of I, P and P_m frames which are arranged alternately. In the predicted (GoP selected) structure, the type of frames are predicted by the GoP selection network, targeting on better bitrate and machine vision performance trade-off.

sures high-quality visuals for human perception, as measured by metrics such as PSNR or MS-SSIM [41]. However, this method introduces significant redundancy, especially since downstream tasks like video object detection or tracking predominantly concentrate on specific regions of interest rather than the entirety of the frame. Consequently, it is crucial for our study to develop an approach that reduces this redundancy in the coding bitstream.

Inspired by Hu et al., who developed a mode prediction technique specifically designed for the human visual system within a slimmable encoder and decoder, we introduce the Dynamic Vision Mode Prediction (DVMP) module. This innovative module autonomously selects the optimal coding mode and decides whether each feature element should be encoded and transmitted, as illustrated in Fig. 3.

To elucidate the functionality of our proposed Dynamic Vision Mode Prediction (DVMP) module, we utilize the encoded residual feature m_t as an illustrative example. Assume m_t possesses dimensions $c \times h \times w$, indicating c channels each with a spatial dimension of $h \times w$. The hyperprior network then predicts the mean and variance for each element, resulting in dimensions of $2c \times h \times w$. The mode prediction network's architecture, detailed in the upper branch, consists of two convolution layers and three ResBlocks. To render the "mode prediction" process differentiable, we employ the Gumbel Softmax technique during training to ascertain the skip mode for each encoded residual feature element. This module generates a binary mask for the feature map m_t, where a "1" denotes elements imperative for machine vision that require entropy coding, and a "0" signifies elements either irrelevant for machine vision or accurately predictable by the hyperprior network, thus streamlining the entropy coding process and bitrate reduction. Note that the showed architecture is not suitble for prior models with autoregressive components, like which in DCVC [18]. However, we can extend the module into an autoregressive way to enable the support for DCVC. The detailed structure is shown in the supplementary material.

To optimize the DVMP module, the loss function can be given by:

$$\mathcal{L}_m = R + \lambda_m \mathcal{L}_t \tag{1}$$

where R denotes the coding bitrate and $\mathcal{L}_t$ denotes the loss of downstream vision task, respectively. λ_m is a hyperparameter used to control the trade-off.

3.3. DivGoP → GoP Selection

Obtaining the P_m frames, we firstly use a hand-crafted structure in a GoP, where the P_m frames and P frames are arranged alternately (referred as "DivGoP" and shown in Fig.4). It is observed that this arrangement can bring a certan degree of bitrate saving in detection, since the P_m frames cost much less bitrate than P frames while maintaining the motion and machine vision information. Experimental findings, as depicted in the "FVC DivGoP" curve within Fig. 5, demonstrate that the "$I, P_m, P, \cdots$" structure markedly achieve about 18.6% bitrate saving.

The "DivGoP" configuration, while methodically structured, may not be universally optimal due to the variability in motion and object information across different videos, which could necessitate distinct GoP structures. For instance, videos characterized by rapid target movement might demand a higher frequency of P frames to preserve the quality of reconstruction, as an excess of P_m frames may lead to degradation that negatively affects subsequent frames. Conversely, in scenarios where the camera remains static or the target movement is minimal, the introduction of more P_m frames could be feasible without significantly impacting the reconstruction quality of following frames. Stimulated by this notion, our research endeavors to tailor the GoP structure dynamically, enhancing codec performance for machine vision tasks. To approach this issue, we adopt FVC [14] as the foundational codec, select a GoP size of 10, and address the detection task utilizing the MOT Dataset. The optimization problem is thus defined:

$$\arg \min_{\theta} \quad R(\theta) + \lambda_p L_{\det}(\theta) \tag{2}$$

where R denotes the bitrate, L_{det} denotes the detection loss, λ_p is a weighting factor that balances the effects of bitrate and detection loss, and θ denotes the setting of GoP structure here.

Obtaining a GoP, each of the 9 predicted frames can be either P frame or P_m frame, which can be dynamically decided. To explore the optimal GoP structure for target function 2, and evaluate the performance gap between DivGoP and optimal GoP structure, we use a deep-first-searching (DFS) algorithm, where we use different λ for different bitrate models. The DFS algorithm searches for the structure that makes the target function smallest for each GoP. Results are shown in the "FVC DFS" curve of Fig.5. The application of the DFS algorithm led to a substantial increase in Bpp-mAP performance,, achieving about 32.3% bitrate saving in terms of the BD-RATE metric, while the "Div-GoP" method achieves about 18.6% bitrate saving. Similar conclusions can also be observed on DCVC [18] and TCM [29], reinforcing our hypothesis that distinct videos necessitate unique GoP structures tailored for machine vision tasks. However, the DFS algorithm is extreamly time consuming, and the labels do not exist during real inference process, making the DFS not avaliable to be applied in real time encoding. How to further find an "adaptive GoP structure" in an available time complexity is worth exploring.

From this perspective, we introduce our GoP selection method to adaptively predict the GoP structure at encoding stage, which has a low time complexity and is available in application. Illustrated in Fig.2 (c), our GoP selection network operates in two phases: the pre-analysis and the GoP prediction stages. Initially, input frames undergo pre-analysis, where a detector identifies objects and generates an overlapping mask for each frame, while the RAFT [34] optical flow estimation method calculates optical flow features for each predicted frame using its adjacent predecessor. These features are then refined by masking with the object bounding box mask to produce masked optical flow features, alongside motion feature priors (mean and variance, averaged by channel) derived from the codec's motion encoder. Subsequently, these processed features are input into the GoP prediction network, which employs a feature extractor with convolution blocks for downsampling and aggregation, followed by adaptive average pooling and linear layers that output a normalized s_{logit} for this GoP, which comprises two elements that sum to one. The averaged detection confidence information from the detector is also injected into the linear layers as meta data.

Upon acquiring the s_{logit}, we employ the Gumbel softmax sampling technique throughout the training phase to iteratively generate a binary value for each frame. This process culminates in the formation of a GoP structure vector, where the binary value—0 for a P_m frame and 1 for a P frame—specifies the type of each frame within the GoP.
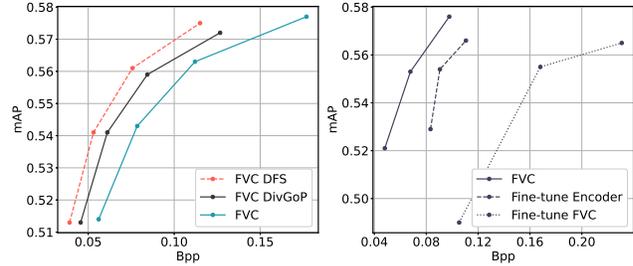

Figure 5. Left: Using DFS to search for the near-optimal GoP structure for Bpp-mAP. Right: Results of simply fine-tuning FVC

During inference stage, we apply a softmax operation on s_{logit} to derive probabilities, which are then used to evenly distribute 1s and 0s, thereby determining the actual encoded GoP structure. Similar to our mode prediction network, the loss function for optimizing GoP structure selection network is given by:

$$\mathcal{L}_g = \bar{R} + \lambda_g \bar{\mathcal{L}}_t \tag{3}$$

where the $\bar{R}$ and $\bar{\mathcal{L}}_t$ denote the GoP-wise average coding bitrate and average loss of downstream vision task.

4. Experiments

In this section, we describe the dataset settings, downstream task models, and evaluation metrics. We then provide empirical evidence underscoring the effectiveness of our proposed method across multiple tasks and codecs. Lastly, we validate the contributions of the mode prediction and GoP selection modules to performance improvements through ablation studies.

Datasets. For multi-object tracking (MOT) task, we choose MOT17 dataset. For video object detection (VOD) task, we adopt the Imagenet VID dataset [28] for training and evaluation. For video action recognition (VAR) task, we adopt UCF101[32] dataset. For the data processing procedure, we downsample the short side of the evaluation videos to 512 for MOT17 dataset and keep the original frame ratios for both training and evaluation. For the evaluation of VOD, we follow the most VOD works to resize the frames to 512×512 for training and 576×576 for evaluation.

Downstream Task Models. For MOT task, we adopt the popular ByteTrack [46] method as the backbone model. The model weights are provided by their official framework. For VOD task, we adopt the detector YOLOV [30], and the model weights are also provided by its official framework. For the VAR task, we adopt the TSM [22] model as the recognition network, the weights are provided by mmaction2 [5]. During training and evaluation of our proposed framework, the weights of the downstream task models are fixed, without any fine-tuning.
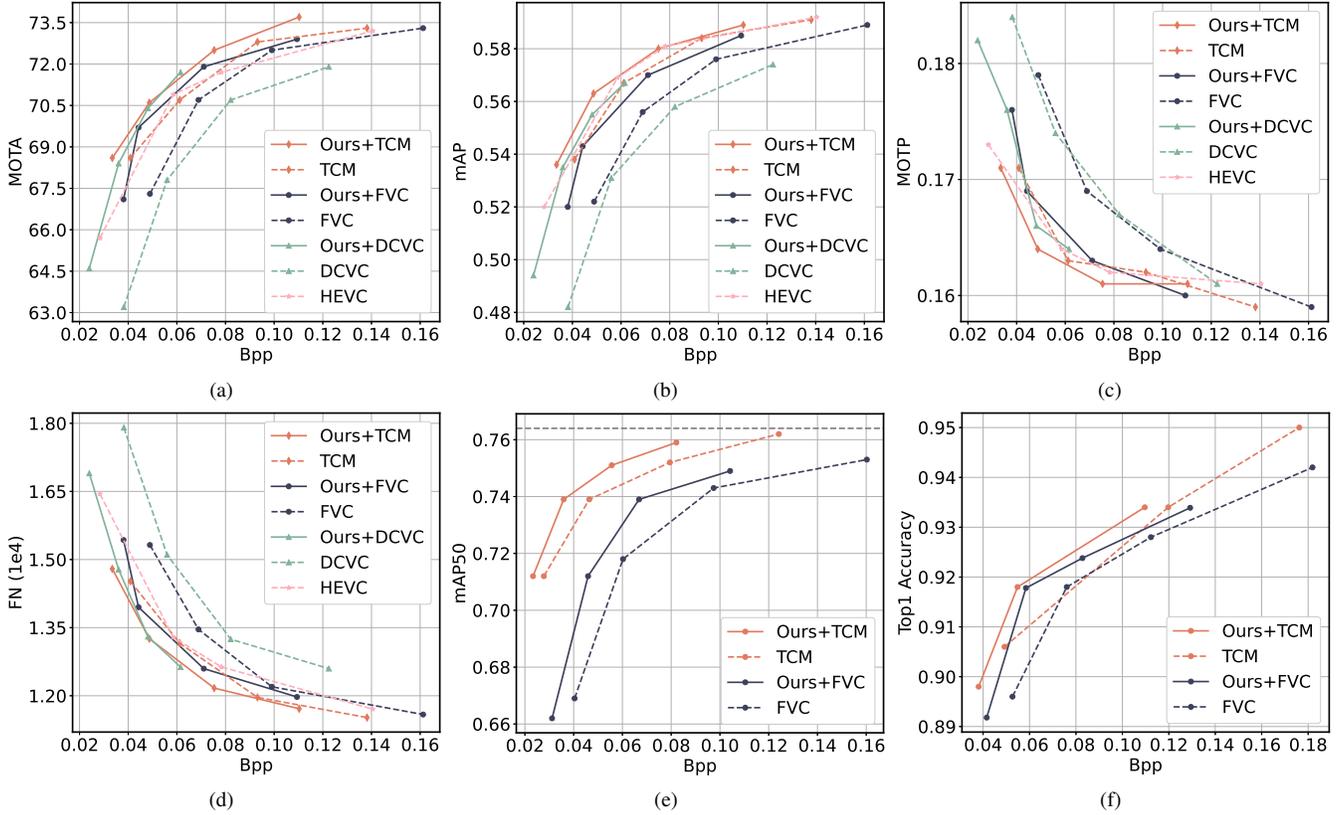
Figure 6. (a) Bpp-MOTA curves on MOT Dataset. (b) Bpp-mAP curves on MOT Dataset. (c) Bpp-MOTP curves on MOT Dataset. (d) Bpp-FN curves on MOT Dataset. (e) Bpp-mAP50 curves on ImageNet VID Dataset using YOLOV [30]. (f) Bpp-Top1 Accuracy curves on UCF101 Dataset.

Evaluation Metrics. For compression, We use bpp (bits per pixel) to measure the average number of bits for one pixel. For the MOT task, we adopt the MOTA (multiple object tracking accuracy), MOTP (multiple object tracking precision), FN (false negative detection number) and mAP (mean average precision) metrics to evaluate the tracking and detection performance. For the VOD task, we adopt the mAP50 metric to evaluate the detection performance, following most VOD methods. For the VAR task, we adopt the Top1 Accuracy as the performance indicator.

Implementation Details. Our whole framework is implemented on PyTorch with CUDA support and trained on eight V100 or A100 GPUs. We use the Adam optimizer by setting the learning rate as 0.0001, β_1 as 0.9 and β_2 as 0.999, respectively. The whole training process has the follow stages: Firstly, we use the loss function 1 to train the mode prediction network and obtain the P_m frames. This stage we use single frame training. Note that for TCM, DVMP is applied to both motion and contextual information. For DCVC and FVC, DVMP is exclusively utilized for motion information, and all residual/contextual information is omitted. After that, we introduce the GoP selection network in a GOP length unit. Using the GoP size of 32

on MOT and VAR tasks, we devide the GoP into 2 mini GoPs with size of 16, and the predicted vector of each mini GoP has a length of 15. For VOD task, we directly set the GoP size as 16. Using multi-frame training and loss function 3, we fix the weights of mode prediction network and train GoP selection network for about 100 epochs on MOT half-train dataset and 20 epochs on ImageNet VID training dataset, respectively for multi-object tracking and video object detection tasks. The whole training procedure takes about 2 days for each task.

4.1. Multi-Object Tracking

In Fig. 6, we present a comparison between our method and various codecs, specifically DCVC, FVC, and TCM. It is evident that our approach yields a markedly improved trade-off in terms of Bpp-MOTA, Bpp-mAP, Bpp-MOTP and Bpp-FN metrics. When compared to the baseline codecs FVC, DCVC and TCM, our control method results in bitrate savings of approximately 27.54%, 41.82% and 12.64%, respectively, for equivalent MOTA metrics. With DCVC established as the anchor baseline, Table 1 displays the BD-RATE values, where our controlled TCM, controlled DCVC and controlled FVC codecs demonstrate the best, second-
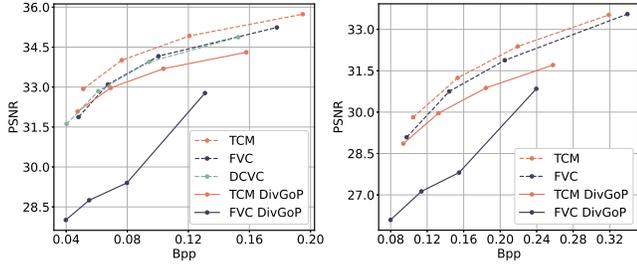
Figure 7. Video Reconstruction Quality in terms of Bpp-PSNR on HEVC Class B dataset (Left) and Class C dataset (Right)

Figure 8. Left: Ablation study for Mode Prediction network. Right: Ablation study for GoP Selection network

Table 1. BD-RATE results on multi-object tracking.

Method	MOTA	mAP	mAP50	MOTP	FN	Average
DCVC [18]	0.0	0.0	0.0	0.0	0.0	0.0
Ours+DCVC	-41.82	-39.43	-40.60	-37.73	-41.74	**-40.26**
FVC [14]	-5.32	-9.75	-14.53	-1.39	-6.20	-7.44
Ours+FVC	-34.28	-31.27	-32.09	-35.02	-32.89	**-33.11**
TCM [29]	-25.19	-32.34	-31.02	-39.85	-26.44	-30.97
Ours+TCM	-40.82	-45.10	-46.15	-51.66	-38.98	**-44.54**
HEVC [33]	-33.88	-31.35	-40.43	15.80	-34.02	-24.78

best and third-best performance, respectively. Additionally, we contrast our method with the task-oriented "One-to-one Codecs" of VCM, with detailed results provided in the supplementary materials.

4.2. Video Object Detection

For the evaluation of the video object detection task, we juxtapose our method with established codecs such as FVC, and TCM, with the comparative results illustrated in Fig. 6(e). Our method is noted to achieve a superior trade-off, particularly in the Bpp-mAP50 metric, when applied to the downstream model YOLOV. For instance, in comparison to the FVC and TCM baseline codecs, our method facilitates bitrate reductions of approximately 21.76% and 20.42%, respectively, while maintaining the same mAP50 metric. The uncompressed video represents the upper bound for the mAP50 metric, recorded at 76.4, and is delineated by the grey line within the figure.

4.3. Video Action Recognition

We conducted an evaluation of our method on the UCF101 dataset, benchmarking against TCM and FVC codecs. The outcomes of this comparison are depicted in Fig. 6(f). For the task of video action recognition, rather than initiating a new set of network trainings, we employed pre-trained mode prediction networks previously utilized for the MOT task and use the "DivGoP" structure for validation. The original videos were compressed using the various codecs, and the resulting decoded videos served as inputs to the TSM model as cited in Lin et al. [22]. The experimental results indicate that our method maintains efficacy in this
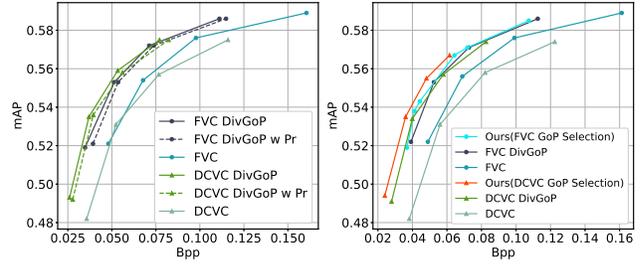
context, securing approximately 21.81% and 26.03% in bitrate savings for TCM and FVC, respectively.

4.4. Video Reconstruction Quality

When it comes to the performance of reconstructing the video for human viewing, the situation is straightforward for conventional video codecs, no matter learned or traditional ones: just encode the frames, transmit and decode the compressed bitstream, and measure their quality using appropriate metrics (PSNR, MS-SSIM, etc). In our method, there're two options: Control the encoding procedure for human viewing, which is the ideal way for best frame reconstruction quality, or just continue the encoding procedure used for machine vision tasks.

Note that our method does not change the weights of the original encoder and decoder in pre-trained DVC, so using this option can directly recover the human viewing quality to original best performance of the codecs. On the other hand, we also evaluate the human viewing quality of the encoded bitstream for machine, using "DivGop" encoding structure on HEVC Class B and Class C test datasets. These are referred to as "FVC DivGoP" and "TCM DivGoP" in Fig.7. Results show that our control for machine does decrease the video reconstruction quality of the original TCM codec by about 1dB in PSNR. Besides, it is also observed that the PSNR of residual codec FVC drops more sharply than that of conditional codecs like TCM. However, these are acceptable since we can always adjust the encoder to compress the video for human viewing when it's needed.

4.5. Ablation Study

Mode Prediction. In our method, we propose mode prediction network to decide the coding modes of motion and residual/contextual feature elements for machine vision, mainly targeting on reducing the bitrate and keep the crucial information for important objects in the frame. Moreover, mainstream DVC, including FVC, DCVC, TCM and others, use motion estimation and warping operation to generate a rough predicted frame, which we call P_r frame here. The P_r frame also meets our requirement of "consuming

| Input image | Contextual skip mode mask | Masked motion feature | Masked Contextual feature |

Figure 9. Visualization results of the mode prediction networks.

Table 2. Encoding Latency vs. Performance.

Method	Encoding Latency(s)	BD-BR Performance
DCVC	1.675	0.0
Ours+DCVC	1.738	**-40.26**
FVC	0.116	-7.44
Ours+FVC	0.159	**-33.11**

lower code streams but retaining object and motion information" to a certain extent. So comparing it with the P_m frame used in our method is also a point worthy of attention. In "DivGoP" structure, we use the P_r frame to replace the original P_m frames, and the GoP structure will be $[I, P_r, P, P_r, ...]$. As shown in the left figure in Fig.8, based on DCVC and FVC, we compare the structure using P_r frames with the original "DivGoP" structure using P_m frames, corresponding to the "DCVC/FVC DivGoP w P_r" and "DCVC/FVC DivGoP" curves respectively. It is observed that the "DivGoP w P_r" curve also achieves about 28.68% and 16.24% bitrate savings for DCVC and FVC, respectively. But our "DivGoP" curve shows better trade-off, by achieving 32.67% and 19.43% bitrate savings.

GoP Selection. We prove the effectiveness of Gop Selection module in terms of rate-precision performance on MOT dataset. As shown in the right figure in Fig. 8, we compare the Bpp-mAP performance of "DCVC/FVC Gop Selection" and "DCVC/FVC DivGoP" structure on MOT Datast. Using DCVC as an example, experiment results show that our Gop Selection achieves about 39.43% bitrate saving, which does show better trade-off than that 32.67% bitrate saving of hand-crafted "DivGoP" structure.

Simply fine-tuning DVC for machine vision tasks. We employed the loss function $L_f = R + \lambda_1 MSE + \lambda_2 L_{det}$ to fine-tune the FVC encoder and the entire FVC codec for the MOT task. As shown in Fig. 5 right, while simple fine-tuning improves accuracy, it also notably increases the bitrate. This phenomenon could be attributed to the degradation in the fidelity of decoded frames when optimized for downstream tasks. Due to the inter-frame reference nature of video codecs, this degradation is propagated across frames, resulting in a high bitrate consumption. This observation aligns with findings reported in DeepSVC [21], indicating that such task-specific performance can not be achieved with simply finetuning a DVC. This also reflects the superiority of our method.

Visualization Results. We present the visualization results of our mode prediction network, as shown in Fig. 9. The first column shows the input images, the second column shows the skip mode masks for contextual features (averaged by channel), and the third and fourth columns are

masked motion and contextual features, respectively, where the red color indicates large value and green color indicates small value. It is observed that the predicted coding mode mask for contextual features kept the regions of the small objects in the frames, and masked the conspicuous objects which are easy to recognize and the background regions. This shows the network's preference for objects in the frame. As for the motion information, it is observed that the predicted masks kept most of the elments of moving objects, which are crucial for detection and tracking.

Complexity and Encoding Latency Analysis. The parameter numbers of our dynamic vision mode prediction network and GoP selection network are 1.48M and 5.36M. Since our method controls the video encoder during the encoding phase, such an operation brings encoding latency. To measure the impact on the encoding phase, we include a comparison of average encoding latency versus performance on MOT Dataset (960×512) in Tab. 2. While our method bring an increase in encoding latency (1.738s vs. 1.675s), the improvement in R-P performance is substantial. Besides, the decoding procedure and decoder side models are not changed, which brings convenience to the deployment of the decoder and support for multiple tasks.

5. Conclusion

In this paper, we propose a flexible framework for deep video compression models, where the pre-trained encoders can be controlled to change the encoding pipeline for machine vision tasks, achieving significant better bpp-precision trade-off without changing the decoders or decoding procedure. The controlling methods are attributed to two main developments, including dynamic vision mode prediction network and GoP structure selection network. Experiment results show that our framework is general for residual and conditional deep video codecs and different downsteam vision tasks like detection, tracking and action recognition. Our comprehensive evaluation, conducted against existing deep video codecs, confirms the superior performance of our models, establishing a new benchmark for future deep video coding for machine. The effectiveness of each component is clearly evidenced through extensive ablation studies.

References

[1] Mohammad Akbari, Jie Liang, and Jingning Han. Dsslic: Deep semantic segmentation-based layered image compression. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2042–2046. IEEE, 2019. 1, 2

[2] Benjamin Bross, Ye-Kui Wang, Yan Ye, Shan Liu, Jianle Chen, Gary J. Sullivan, and Jens-Rainer Ohm. Overview of the versatile video coding (vvc) standard and its applications. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(10):3736–3764, 2021. 1, 2

[3] Zhuo Chen, Kui Fan, Shiqi Wang, Lingyu Duan, Weisi Lin, and Alex Chichung Kot. Toward intelligent sensing: Intermediate deep feature compression. *IEEE Transactions on Image Processing*, 29:2230–2243, 2020. 2

[4] Hyomin Choi and Ivan V Bajić. Scalable image coding for humans and machines. *IEEE Transactions on Image Processing*, 31:2739–2754, 2022. 1, 2

[5] MMAction2 Contributors. Openmmlab's next generation video understanding toolbox and benchmark. https://github.com/open-mmlab/mmaction2, 2020. 5

[6] Yan Dai, Ziyu Hu, Shuqi Zhang, and Lianjun Liu. A survey of detection-based video multi-object tracking. *Displays*, page 102317, 2022. 3

[7] Ling-Yu Duan, Tiejun Huang, and Wen Gao. Overview of the mpeg cdvs standard. In *2015 Data Compression Conference*, pages 323–332, 2015. 2

[8] Ling-Yu Duan, Yihang Lou, Yan Bai, Tiejun Huang, Wen Gao, Vijay Chandrasekhar, Jie Lin, Shiqi Wang, and Alex Chichung Kot. Compact descriptors for video analysis: The emerging mpeg standard. *IEEE MultiMedia*, 26(2): 44–54, 2019. 2

[9] Yann Dubois, Benjamin Bloem-Reddy, Karen Ullrich, and Chris J Maddison. Lossy compression for lossless prediction. *Advances in Neural Information Processing Systems*, 34:14014–14028, 2021. 2

[10] Ruoyu Feng, Xin Jin, Zongyu Guo, Runsen Feng, Yixin Gao, Tianyu He, Zhizheng Zhang, Simeng Sun, and Zhibo Chen. Image coding for machines with omnipotent feature learning. In *European Conference on Computer Vision*, pages 510–528. Springer, 2022. 3

[11] Kristian Fischer, Fabian Brand, and André Kaup. Boosting neural image compression for machines using latent space masking. *arXiv preprint arXiv:2112.08168*, 2021. 1, 2

[12] Hadi Hadizadeh and Ivan V Bajić. Learned scalable video coding for humans and machines. *arXiv preprint arXiv:2307.08978*, 2023.

[13] Yueyu Hu, Shuai Yang, Wenhan Yang, Ling-Yu Duan, and Jiaying Liu. Towards coding for human and machine vision: A scalable image coding approach. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2020. 1, 2

[14] Zhihao Hu, Guo Lu, and Dong Xu. Fvc: A new framework towards deep video compression in feature space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1502–1511, 2021. 1, 2, 4, 7

[15] Zhihao Hu, Guo Lu, Jinyang Guo, Shan Liu, Wei Jiang, and Dong Xu. Coarse-to-fine deep video coding with hyperprior-guided mode prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5921–5930, 2022. 1

[16] Sayed Hossein Khatoonabadi and Ivan V Bajic. Video object tracking in the compressed domain using spatio-temporal markov random fields. *IEEE transactions on image processing*, 22(1):300–313, 2012. 3

[17] Congcong Li, Xinyao Wang, Longyin Wen, Dexiang Hong, Tiejian Luo, and Libo Zhang. End-to-end compressed video representation learning for generic event boundary detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13967–13976, 2022. 3

[18] Jiahao Li, Bin Li, and Yan Lu. Deep contextual video compression. *Advances in Neural Information Processing Systems*, 34:18114–18125, 2021. 1, 2, 4, 5, 7

[19] Jiahao Li, Bin Li, and Yan Lu. Neural video compression with diverse contexts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22616–22626, 2023. 1

[20] Xin Li, Jun Shi, and Zhibo Chen. Task-driven semantic coding via reinforcement learning. *IEEE Transactions on Image Processing*, 30:6307–6320, 2021. 1, 2

[21] Hongbin Lin, Bolin Chen, Zhichen Zhang, Jielian Lin, Xu Wang, and Tiesong Zhao. Deepsvc: Deep scalable video coding for both machine and human vision. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 9205–9214, 2023. 3, 8

[22] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7083–7093, 2019. 5, 7

[23] Qiankun Liu, Bin Liu, Yue Wu, Weihai Li, and Nenghai Yu. Real-time online multi-object tracking in compressed domain. *arXiv preprint arXiv:2204.02081*, 2022. 3

[24] Guo Lu, Wanli Ouyang, Dong Xu, Xiaoyun Zhang, Chunlei Cai, and Zhiyong Gao. Dvc: An end-to-end deep video compression framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11006–11015, 2019. 1, 2

[25] Guo Lu, Xiaoyun Zhang, Wanli Ouyang, Li Chen, Zhiyong Gao, and Dong Xu. An end-to-end learning framework for video compression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10):3292–3308, 2021. 1, 2

[26] Guo Lu, Xingtong Ge, Tianxiong Zhong, Jing Geng, and Qiang Hu. Preprocessing enhanced image compression for machine vision. *arXiv preprint arXiv:2206.05650*, 2022. 2

[27] Fabian Mentzer, Eirikur Agustsson, Johannes Ballé, David Minnen, Nick Johnston, and George Toderici. Neural video compression using gans for detail synthesis and propagation. In *European Conference on Computer Vision*, pages 562–578. Springer, 2022. 2

[28] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large

scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015. 5

[29] Xihua Sheng, Jiahao Li, Bin Li, Li Li, Dong Liu, and Yan Lu. Temporal context mining for learned video compression. *IEEE Transactions on Multimedia*, 2022. 1, 2, 5, 7

[30] Yuheng Shi, Naiyan Wang, and Xiaojie Guo. Yolov: making still image object detectors great at video object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2254–2262, 2023. 5, 6

[31] Zheng Shou, Xudong Lin, Yannis Kalantidis, Laura Sevilla-Lara, Marcus Rohrbach, Shih-Fu Chang, and Zhicheng Yan. Dmc-net: Generating discriminative motion cues for fast compressed video action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1268–1277, 2019. 3

[32] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 5

[33] Gary J. Sullivan, Jens-Rainer Ohm, Woo-Jin Han, and Thomas Wiegand. Overview of the high efficiency video coding (hevc) standard. *IEEE Transactions on Circuits and Systems for Video Technology*, 22(12):1649–1668, 2012. 1, 2, 7

[34] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020. 5

[35] Yuan Tian, Guo Lu, Xiongkuo Min, Zhaohui Che, Guangtao Zhai, Guodong Guo, and Zhiyong Gao. Self-conditioned probabilistic learning of video rescaling. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4490–4499, 2021. 2

[36] Yuan Tian, Guo Lu, Guangtao Zhai, and Zhiyong Gao. Non-semantics suppressed mask learning for unsupervised video semantic compression. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13610–13622, 2023. 1, 3

[37] Yuan Tian, Guo Lu, Yichao Yan, Guangtao Zhai, Li Chen, and Zhiyong Gao. A coding framework and benchmark towards low-bitrate video understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 3

[38] Robert Torfason, Fabian Mentzer, Eirikur Agustsson, Michael Tschannen, Radu Timofte, and Luc Van Gool. Towards image understanding from deep compression without decoding. *arXiv preprint arXiv:1803.06131*, 2018. 1, 2

[39] Shiyao Wang, Hongchao Lu, and Zhidong Deng. Fast object detection in compressed video. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7104–7113, 2019. 3

[40] Shurun Wang, Shiqi Wang, Wenhan Yang, Xinfeng Zhang, Shanshe Wang, Siwei Ma, and Wen Gao. Towards analysis-friendly face representation with scalable feature and texture compression. *IEEE Transactions on Multimedia*, 2021. 1, 2

[41] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, pages 1398–1402. Ieee, 2003. 2, 4

[42] T. Wiegand, G.J. Sullivan, G. Bjontegaard, and A. Luthra. Overview of the h.264/avc video coding standard. *IEEE Transactions on Circuits and Systems for Video Technology*, 13(7):560–576, 2003. 1, 2

[43] Chao-Yuan Wu, Manzil Zaheer, Hexiang Hu, R Manmatha, Alexander J Smola, and Philipp Krähenbühl. Compressed video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6026–6035, 2018. 3

[44] Ning Yan, Changsheng Gao, Dong Liu, Houqiang Li, Li Li, and Feng Wu. Sssic: semantics-to-signal scalable image coding with learned structural representations. *IEEE Transactions on Image Processing*, 30:8939–8954, 2021. 1, 2

[45] Ren Yang, Luc Van Gool, and Radu Timofte. Perceptual learned video compression with recurrent conditional gan. *arXiv preprint arXiv:2109.03082*, 1, 2021. 2

[46] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. In *European Conference on Computer Vision*, pages 1–21. Springer, 2022. 5

[47] Ezgi Özyılkan, Mateen Ulhaq, Hyomin Choi, and Fabien Racapé. Learned disentangled latent representations for scalable image coding for humans and machines. In *2023 Data Compression Conference (DCC)*, pages 42–51, 2023. 1, 2