Non-Semantics Suppressed Mask Learning for Unsupervised Video Semantic Compression

Yuan Tian[1] Guo Lu[1]✉ Guangtao Zhai[1]✉ Zhiyong Gao[1]
[1]Institute of Image Communication and Network Engineering, Shanghai Jiao Tong Unversity
{ee_tianyuan, luguo2014, zhaiguangtao, zhiyong.gao}@sjtu.edu.cn

Abstract

Most video compression methods aim to improve the decoded video visual quality, instead of particularly guaranteeing the semantic-completeness, which deteriorates downstream video analysis tasks, e.g., action recognition. In this paper, we focus on a novel unsupervised video semantic compression problem, where video semantics is compressed in a downstream task-agnostic manner. To tackle this problem, we first propose a Semantic-Mining-then-Compensation (SMC) framework to enhance the plain video codec with powerful semantic coding capability. Then, we optimize the framework with only unlabeled video data, by masking out a proportion of the compressed video and reconstructing the masked regions of the original video, which is inspired by recent masked image modeling (MIM) methods. Although the MIM scheme learns generalizable semantic features, its inner generative learning paradigm may also facilitate the coding framework memorizing non-semantic information with extra bit costs. To suppress this deficiency, we explicitly decrease the non-semantic information entropy of the decoded video features, by formulating it as a parametrized Gaussian Mixture Model conditioned on the mined video semantics. Comprehensive experimental results demonstrate the proposed approach shows remarkable superiority over previous traditional, learnable, and perceptual quality-oriented video codecs, on three video analysis tasks and seven datasets.

1. Introduction

Video compression has been actively researched over the past few decades. Most methods, including traditional methods [102][14] and learnable ones [83][61], aim at improving the reconstructed video quality for human perception, rather than particularly preserving AI task-required semantic information, *e.g.*, human key-points and object shapes. This degrades downstream AI tasks [135][104].

To tackle this problem, lots of research efforts have been devoted to a new problem of Video Coding for Machine (VCM), *i.e.*, lossy video compression for supporting downstream AI tasks. For example, early works [142][23][8] and standards [46][44][45] additionally transport the manually-designed image descriptors, supporting limited tasks with undesirable performance. Later, some methods [86][53][33] improve the traditional codec with hand-crafted designs to better cope with the specific tasks, *e.g.*, saliency-aware bit allocation for object detection task [53]. Meanwhile, some methods [34][29][28][100][51] compress the feature maps of AI models instead of the images, where the tail task modules shall adapt to the features by supervised learning. Besides, most scalable coding methods [79][133][126][35][36][103] are optimized by supervised learning or task-relevant feature matching loss functions. Despite these efforts, rare methods are task-agnostic while still exploiting data-driven semantics. Task-agnostic decouples the coding system from downstream tasks and is friendly to data scarcity scenarios. Data-driven objective prompts the system to learn a more generalizable semantic representation than hand-crafted designs.

In this paper, we focus on a novel Unsupervised Video Semantic Compression task that satisfies the two above requirements, where compressed videos readily support various analysis tasks. Considering plain video codecs are already of powerful visual coding capability, it is natural to build the semantic compression framework upon them for inheriting the advantages. However, it is non-trivial to deal with the semantic information lost during the compression, especially without the guidance of task-specific data labels.

To address these issues, we first propose a simple yet effective Semantic-Mining-then-Compensation (SMC) framework as a baseline method, to improve current plain video codecs with better semantic coding capability. Specifically, the semantic feature of the original video and the lossy video is extracted by neural networks on the encoder side, and only the residual part is transported. On the decoder side, the lossy video and its semantics compensated by the residual, together synthesize the final decoded video. As for the self-supervised optimization of the framework, we mask out a large proportion of the decoded video patches, and use the unmasked parts to reconstruct the masked regions of the original video, inspired by recent Masked Image Modeling (MIM) methods [56][113]. The reconstruction task facilitates decoded videos to be of rich semantics in terms of both intra-patch appearance and inter-patch interactions. Although the MIM scheme learns generalizable semantic features, its inner generative learning paradigm also facilitates the coding framework memorizing non-semantic information [41], which makes the extracted semantic features contain redundant information,

✉ Corresponding Author

consuming extra bitcost. To suppress this deficiency, we explicitly decrease the non-semantic information entropy of the video, by formulating it as a parameterized Gaussian Mixture Model conditioned on the mined video semantics. The alternative semantic learning and non-semantic suppressing procedures make the system bootstrapping itself toward more efficient semantic coding. As a result, it shows remarkable performance on a wide range of tasks without leveraging any data labels.

Contributions:

- We focus on a novel unsupervised video semantic compression problem, proposing a concise yet effective baseline framework dubbed SMC.
- Our work is the first one that applies Masked Image Modeling (MIM) scheme to semantic coding problem, aiming to learn a semantic representation that is applicable to various downstream tasks.
- We propose the Non-Semantics Suppressed (NSS) learning strategy to better adapt the general MIM scheme to the compression problem, suppressing the framework from encoding non-semantic information.
- Our approach demonstrates notable superiority over previous traditional, learnable and perceptual codecs, on three video analysis tasks and seven datasets.

2. Related Works

Video Compression. Previous video codecs, including traditional ones [121][102][14][138], learnable ones [83][61][73][85][30][84][82][83] and mixed ones [106][110][139], are designed to achieve better pixel-wise signal quality metrics, *e.g.*, PSNR and MS-SSIM [118], which mainly serve the human visual experience. Recently, there are also some generative video coding methods [132][88] that mainly consider visual comfort and perceptual quality [136][145][146][147][55][78].

Video Coding for Machine (VCM). Early standards such as CDVA [45] and CDVS [44][46] propose to pre-extract and transport the image keypoints, supporting image indexing or retrieval tasks. Some works [142][23][8][34][29][28][100][51] compress the intermediate feature maps instead of images. Besides, some works [86][53][33][64] [37][15][140] improve traditional codecs by introducing downstream task-guided rate-distortion optimization strategy or another task-specific feature encoding stream [115][23]. Also, some methods [71][4][134][43][36] optimize the learnable codecs by directly incorporating the downstream task loss. Recently, some methods exploit hand-crafted structure maps [59][47] for semantic coding. Nevertheless, most above methods rely on task-specific labels or hand-crafted/heuristic priors, the effectiveness of which is limited to the targeted tasks.

Recently, some works leverage self-supervised representation learning methods for learning a compact semantic representation. As an pioneering work, Dubois *et al*. [48] theoretically reveals that the distortion term of the lossy rate-distortion trade-off for image classification can be approximated by a contrastive learning objective [57]. However, the compressed semantics are empirically effective to a group of tasks that share the similar prior (*i.e.* , labels are invariant to data augmentations), but may severely discard the semantics useful to other tasks such as detection and segmentation. Recently, Feng *et al*. [51] proposes to learn a unified feature representation for AI tasks from unlabeled data in a similar manner. In these two methods, the downstream models are required to be fine-tuned for adapting to the features. Very recently, Tian *et al*. [107] propose a self-supervised edge representation as the semantic intermediary to constrain the semantic structure of the video. Although working without a task-specific post-adaptation procedure, the edge representation is still highly hand-crafted.

Scalable Coding and Visual-Semantic Fusion Coding. Scalable coding methods [59][133][79][126][35][37] can achieve excellent compression efficiency when measured with the trained tasks, but usually show undesirable results on the tasks/data out of the training scope, due to the supervised learning paradigm. Visual-semantic fusion coding methods [47][3][62][59][148], *a.k.a*, conceptual coding [19][21][22][20], first extract the structure information and the texture information on the encoding side, and then fuse the two parts into a image on the decoding side. The fused images are readily fed into various task and achieve superior performance even at very low bitrate levels. However, almost all these methods employ a pre-trained network to generate semantic segmentation map [47] or edge map [59] as the semantic stream, not fully discarding the task-specific priors.

Compressed Video Analysis. There are also amounts of works such as [80][116][72][117][137] perform video analysis tasks, such as image recognition [39][128][127], action recognition [123][98][16][50][111][105][108] and multiple object tracking (MOT) [67][68][38], in the compressed video domain. But, these methods focus on developing video analysis models that better leverage the partially decoded video stream, such as the motion vector. In contrast, our work focuses on the coding procedure.

Self-Supervised Semantic Learning. Recent methods can be mainly divided into two catogories, *i.e.*, Contrative Learning (CL) ones [57][27][25][91] and Masked Image Modeling (MIM) ones [144][56][113][41]. CL methods use two augmented views of the same image as the positive pair, and other images as negative samples. The learned semantics relies on the employed augmentation strategy [109], and is usually biased to global semantics. Recently, MIM methods have gain increasing attentions. MIM simply predicts masked patches from unmasked ones, while showing remarkably strong performance in downstream tasks. After the pioneering works, *e.g.*, MAE [56] and Beit [7], amounts of works have been proposed for improving the MIM framework [26][125][41] or the prediction target [40][144][119]. Although MIM methods are superior to CL ones in many aspects, when the masked region reconstruction loss serves as the semantic learning objective of a compression system, it also facilitates the system encoding some non-semantic information, and wasting extra bitcosts. Our work solves this problem by explicitly suppressing the non-semantics information within the MAE feature space.

3. Approach

3.1. Framework Overview

We propose the Semantic-Mining-then-Compensation framework SMC, as shown in Figure 1. Let a high-quality X and its compressed lossy version $\tilde{X}$ by plain video codec such as VVC. On the encoder side, SMC additionally transmits the residual semantic information Res that is lost during the lossy compression procedure. On the decoder side, SMC fuses the compensated semantic feature $\hat{S}$ and the lossy video $\tilde{X}$ to synthesize a high-quality video $\hat{X}$, which can be readily processed for various analysis tasks. The framework subcomponents are detailed as follows.

Semantic Extraction Network (Sem-Net). To transform the videos from RGB space to semantic space, the original input frame and the lossy encoded video $\tilde{X}$ are encoded as the semantic representations S and $\tilde{S}$, respectively. The tensor shape of S and $\hat{S}$ is both $\mathbb{R}^{T\times 512\times \frac{H}{32}\times \frac{W}{32}}$, where T and $H \times W$ represent the temporal length and spatial dimensions of the input video. For producing S, we adopt the ResNet18 [58] network as the Sem-Net, but replacing its first Max-pooling layer with a stride two convolution layer for retaining more information. For producing $\tilde{S}$, we adopt a more lightweight network denoted Sem-Net$_\mathrm{s}$, simply consisting of five convolution layers of stride size two and kernel size three. The weights of the two networks are randomly initialized.

Semantic Residual Coding. The residual semantic feature Res between the original video semantic feature S and lossy one $\tilde{S}$ will be compressed by using an auto encoder-decoder (AED) network. Both the encoder and the decoder networks are composed of three causal temporal convolutions [90], ensuring the current feature only depends on the previous state, which is consistent with the low-delay P frame (LDP) mode of coding methods [131][132]. Adding back the reconstructed residual semantic feature $\hat{Res}$ to $\tilde{S}$, we produce the compensated semantic feature $\hat{S}$.

Semantic-Visual Information Fusion. After obtaining the compensated semantic feature $\hat{S}$ and the lossy video $\tilde{X}$ on the decoder side, we use a UNet-style [96] generator network termed F-Net, where the deep latent features are modulated by $\hat{S}$ with the AdaIN [63] operations, to synthesize the final video $\hat{X}$. $\hat{X}$ will be consumed by machine models.

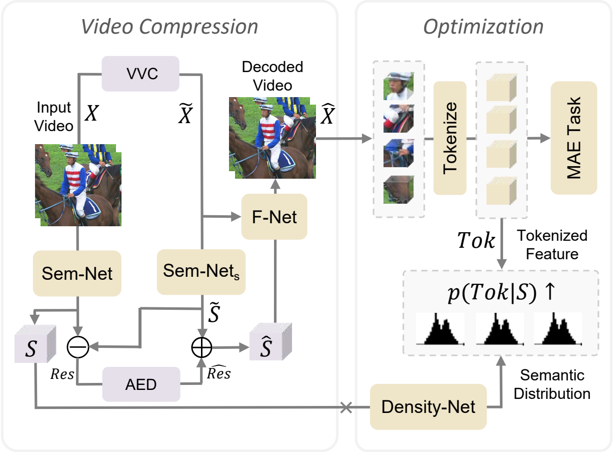

Figure 1: Overview of the proposed SMC framework, which encompasses two bitstreams, *i.e.*, video stream from VVC codec and residual semantic stream encoded by AED network. During the compression, the semantic features of the original video X and the lossy video $\tilde{X}$ are separately mined. Then, the residual semantics Res is transported to compensate the lossy video semantics. Finally, the compensated semantic feature $\hat{S}$ and the lossy video $\tilde{X}$ is fused into a video $\hat{X}$ for supporting various downstream tasks. The framework is optimized via a non-semantics suppressed MAE task loss. × denotes the gradient stopping operation.

3.2. Optimization of Framework

The optimization target is to enforce the decoded video $\hat{X}$ of both rich semantics and good visual quality, while minimizing the bitrate of the residual semantic feature. The whole loss function can be given by $\mathcal{L} = \alpha\mathcal{L}_{sem} + \mathcal{L}_{lpips} + \mathcal{L}_{GAN} + H(Res)$, where $\mathcal{L}_{sem}$ denotes the proposed self-supervised semantic learning loss, which will be detailed in the next section, α is the balancing weight. Following [49], we introduce the combined $\mathcal{L}_{lpips} + \mathcal{L}_{GAN}$ item to regularize the visual quality of $\hat{X}$, where $\mathcal{L}_{lpips}$ denotes the learned image perceptual loss [141], and $\mathcal{L}_{GAN}$ denotes the GAN loss. The discriminator network architecture of the GAN is same as that of PatchGAN [65]. LSGAN loss [87] is adopted for its better stability than vanilla GAN [54].

$H(Res)$ represents the bitrate of the residual semantic stream, which is estimated by the simple bitrate estimation model in [5]. During entropy coding, the latent feature of Res, which is extracted by the encoder part of AED network, will be first quantized and then transformed into the bitstream. Following [6], we approximate the quantization operation by adding the uniform noise in the training stage.

3.3. Self-Supervised Semantic Learning

In this section, we describe how to learn semantic representation from unlabeled videos, which is one of the core challenges for unsupervised semantic coding problem. The semantic learning objective of our framework is based on the MAE framework [56], inspired by the fact that

MAE learns strong and generalizable semantic representation from unlabeled image/video data. Considering the distinct characteristic of the coding problem, *i.e.*, unnecessary information should be excluded from the coding system for saving bitcost, we introduce a Non-Semantics Suppressed (NSS) learning strategy to guide the MAE framework more preserving semantic information within the video.

MAE Learning. Given the decoded video $\hat{X}$, we first divide it into regular non-overlapping patches of size 16×16, and each patch is transformed to tokens by linear embedding, forming the token set Tok. Then a proportion of tokens are randomly masked, and the remaining unmasked ones are fed into the prediction network ϕ (including an encoder and a decoder) for reconstructing the video. The reconstruction loss of the MAE task is given by, $\mathcal{L}_{MAE} = \frac{1}{\mathcal{M}} \sum_{i \in \mathcal{M}} ||\phi[Tok(i)] - X(i)||$, where i is the token index, $\mathcal{M}$ is the set of masked tokens, and X is the ground-truth video. During the optimization procedure, the encoder of ϕ implicitly clusters the video patches/tokens into some semantic centers, and the decoder builds a spatial-temporal reason graph among these semantic primitives to predict the remained region pixels. This facilitates preserving semantics-relevant information within each patch, as well as interactions among different patches. However, with only pixel-wise regularization, the MIM objective also facilitates $\hat{X}$ over-memorizing some non-semantic information, such as the object surface details, which degrades the compression efficiency.

Non-Semantics Suppressed (NSS) Learning. To suppress the non-semantic information leaked from X to $\hat{X}$, we explicitly regularize the information entropy of Tok, conditioned on the mined semantic feature S. However, due to the difficulty of estimating the entropy of a continuous variable, we insert quantization operation in the tail of the tokenization procedure, so that Tok is a discrete variable. Then, we use a Gaussian Mixture Model (GMM) [95] with component number K to approximate its distribution. The distribution of each token $Tok(i)$ is defined by the dynamic mixture weights w_i, means μ_i and log variances σ_i, which are produced by a density parameter estimation network (Density-Net). With these parameters, the distributions can be determined as,

$$p(Tok(i)|S) \sim \sum_{k=1}^{K} w_i^k \cdot \mathcal{N}(\mu_i^k, e^{\sigma_i^k}). \tag{1}$$

Then, the discretized likelihoods of each video patch token can be given by,

$$p(Tok(i)|S) = c(Tok(i) + 0.5) - c(Tok(i) - 0.5), \tag{2}$$

where $c(\cdot)$ is the cumulative function [92] of the GMM in Equation 1. Finally, the non-semantics suppressed mask learning objective can be given by,

$$\mathcal{L}_{Sem} = -\beta \log(p(Tok|S)) + \mathcal{L}_{MAE}, \tag{3}$$

where β is the balancing weight.

Discussion. Although Equation 2 shares a similar format with the bit estimation procedure of the hyper-prior-based image compression methods [6][32], our goal is fundamentally different from them. Our method aims to suppress the extra non-semantic information that is introduced by the MAE task, transporting zero bits, while the method [6] explores the hierarchical redundancies within images, and the estimated bits are additionally transported.

Density-Net. It consists of two convolutions of kernel size three, followed by one temporal causal convolution of temporal kernel size three, aiming to align the semantic feature S to the token feature space. Then, we append three different multiple layer perceptrons (MLPs) for predicting w, μ, and σ in Equation 1. The gradient of S is detached during the back-propagation procedure, forming a self-bootstrapping paradigm, *i.e.*, NSS scheme enforces S capturing semantic-only information, while the semantic-rich S leads to a more principled NSS objective.

4. Experiments

4.1. Evaluation Datasets

For action recognition task, we evaluate it on four large-scale video datasets, UCF101 [101], HMDB51 [70], Kinetics [16], and Diving48 [76]. For multiple object tracking (MOT) task, we evaluate it on MOT17 [89]. For video object segmentation (VOS) task, we evaluate it on DAVIS2017 [94]. We also compare the visual quality of the decoded videos on HEVC Class C dataset [102].

Dataset Processing. During the *training* procedure, we randomly select 60K videos of resolution larger than 1280×720 from the training set of Kinetics400. We downsample the shortest side of the training videos to 256 pixels for removing compression artifacts introduced by prior codecs on YouTube, which follows [122]. During training, the quantization parameter (QP) value of VVC is randomly sampled, so that the framework is learned to adapt to various QPs with a single model. For the evaluations of videos of *action recognition* datasets, we also pre-downsample the shortest side of them to 256 pixels and crop the video to the size 224×224 before the coding procedure. For the evaluation of *MOT*, we adopt the original MOT17 dataset of resolution 1920×1080 because many tracking methods require high-resolution inputs. For the evaluation of *VOS*, we download the high-resolution version of DAVIS2017, which contains videos from 720p to 1080p, and then downsample them to 480p (854×480), which is the input resolution of most VOS methods.

4.2. Experimental Setting

Downstream Task Models. For the *action recognition* task, we adopt the following popular models, *i.e.*, TSM [77], SlowFast [50], and TimeSformer [10], including 2D CNN,

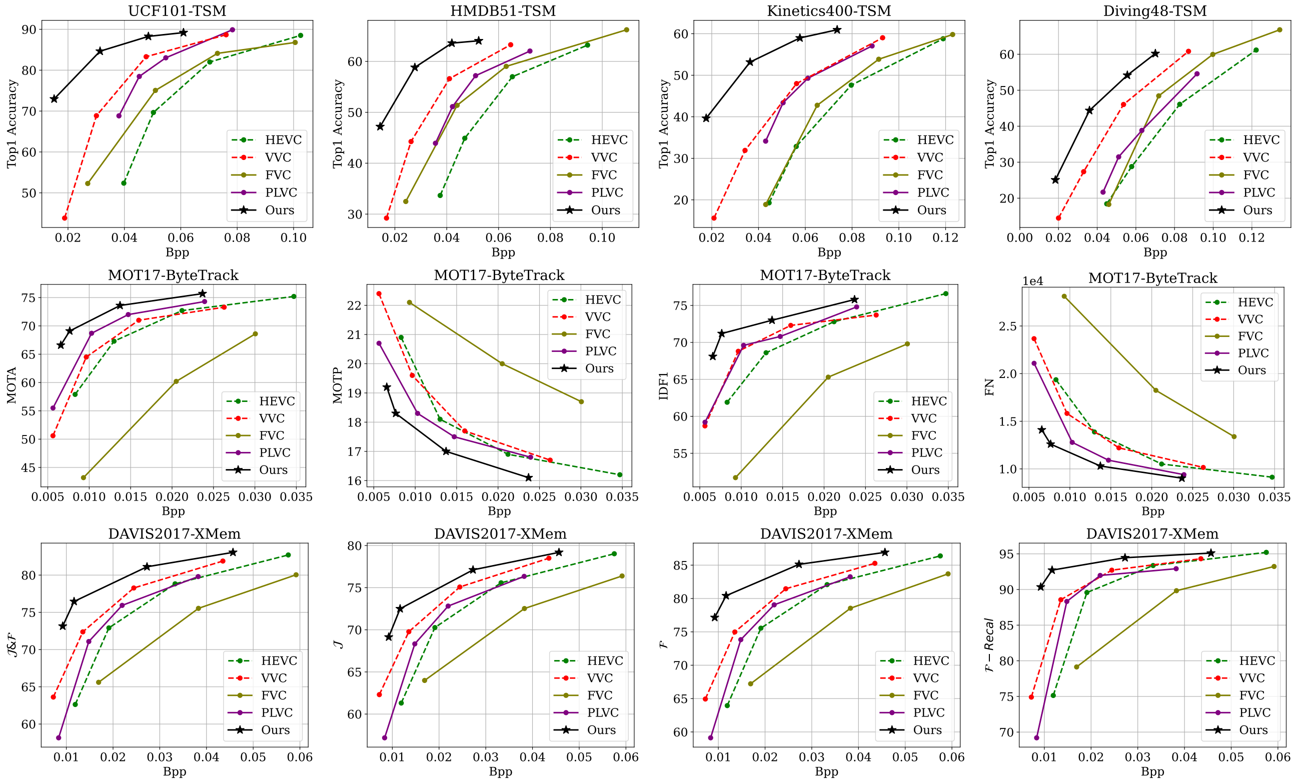

Figure 2: Semantic coding performance on Action Recognition, MOT and VOS tasks. The plot titles are in {Dataset}-{Model} format.

3D CNN and the recent Transformer architectures. The model weights are provided by the MMAction2 framework [2]. For the *MOT* task, we adopt ByteTrack [143], of which the model weights is provided by MMTracking [1]. For the *VOS* task, we adopt XMem [31], of which the model weights is officially released by the authors. We emphasize that we directly feed the decoded videos by our framework into these official models, without any model fine-tuning.

Baseline Codecs. *HEVC* is evaluated with FFmpeg software [112] and x265 codec. *VVC* is evaluated with VVenC1.5.0 software [120], the performance of which is similar to VTM15.0, but consumes much less encoding time. *FVC* [61] is a representative work on learnable codec. *PLVC* [132] is a recent perceptual quality-oriented codec. For traditional codecs, the CRF value is selected from {47,43,39,35}, as setting CRF to larger values does not degrade the downstream task obviously. All codecs and our framework are evaluated on LDP mode with group-of-picture(GOP) size 10, for a fair comparison.

Evaluation Metrics. We use bpp (bit per pixel) to measure the average number of bits used for one pixel in each frame. For the *action recognition* task, we adopt the Top1 accuracy as the performance indicator. For the *MOT* task, we adopt MOTA (multiple object tracking accuracy) [66], MOTP (multiple object tracking precision), FN (false negative detection number) and IDF1, which is the ratio of correctly identified detections over the average number of ground truth and computed detections. For the *VOS* task, the standard metrics Jaccard index $\mathcal{J}$, contour accuracy $\mathcal{F}$ and the average of $\mathcal{J}$ and $\mathcal{F}$ ($\mathcal{J}\&\mathcal{F}$) are adopted. We also report the contour recall $\mathcal{F}$-*Recal*.

Implementation Details. Following VideoMAE [113], the masking ratio for MAE task is set to 90%, where the encoder and decoder networks consist of six and two ViT blocks [42] with divided space-Time attention [10], respectively. The weights of the networks are randomly initialized. K in Equation 1 is empirically set as 5. α and β are set to 1 and 0.1, respectively. We use the Adam optimizer [69] by setting the learning rate as 0.0001, β_1 as 0.9 and β_2 as 0.999, respectively. The resolution of training videos is 256 × 256 and the clip length is 8. The training iteration number is 1000k. The mini-batch size is 24. The system is implemented with Pytorch [93] and it takes about seven days to train the model using eight Nvidia 2080Ti GPUs.

4.3. Experimental Results

Action Recognition. In Table 1, we compare different video compression methods in terms of the action recognition accuracy performance. It is observed that our method outperforms all other methods, including the traditional HEVC and VVC codecs, the learnable FVC codec and the perceptual-oriented codec PLVC by a large margin. When compared with FVC on the UCF101 dataset with TSM performing action recognition, our proposed method achieves a remarkable 22% improvement at 0.04bpp level in terms of Top1 accuracy. Compared with the latest video compression standard VVC on the above setting, our proposed

	TSM Top1 (%)					**Slowfast Top1** (%)			**TimeSformer Top1** (%)	
	UCF101	HMDB51	Kinetics400	Diving48	Diving48	UCF101	HMDB51	HMDB51	UCF101	HMDB51
Bpp	*@0.04*	*@0.04*	*@0.06*	*@ 0.05*	*@ 0.07*	*@0.03*	*@0.02*	*@0.03*	*@0.04*	*@0.04*
HEVC	52.70	36.64	35.28	22.48	37.24	85.13	47.78	62.68	69.92	37.45
FVC [61]	64.61	47.54	37.23	22.94	46.29	79.43	52.15	59.81	72.47	45.11
PLVC [132]	71.40	48.67	48.65	30.13	42.63	87.38	56.33	64.50	70.06	44.59
VVC	76.97	55.72	49.11	42.75	53.23	86.93	59.95	65.74	85.10	57.83
Ours	**86.46**	**62.92**	**59.26**	**51.36**	**60.16**	**88.89**	**63.48**	**67.30**	**89.60**	**60.74**
Original	93.97	72.81	70.73	75.99	75.99	94.92	72.03	72.03	95.43	71.44

Table 1: Results on action recognition. "Original" denotes the performance upper-bound, which is evaluated on the original dataset.

method still surpasses it by 10% at 0.04bpp level. In addition to the largely improved recognition accuracy, we also calculate the bitcost reduction of our method with BDBR algorithm [11] when using VVC as the anchor. Our method substantially saves the bitcost of VVC by 31.65%, 48.93% and 31.41% on UCF101, Kinetics400 and Diving48, respectively. We also provide the rate-performance (RP) curve of different video compression methods on action recognition task. As shown in the first row of Figure 2, our method outperforms all other methods by a large margin on all datasets. Please refer to the supplementary material for the RP curves of Slowfast and TimeSformer model.

Multiple Object Tracking (MOT). In addition to the action recognition task that relies on the global spatial-temporal discriminative semantic cues, we also benchmark the coding methods on a much challenging MOT task. This task requires not only inducing the local location of each object, but also extracting occlusion-robust appearance features for these objects. As shown in Table 2, our method still achieves the best performance in terms of all metrics when compared to all baseline codecs. Our method outperforms FVC, PLVC, VVC codecs by 26.50%, 2.97% and 5.98%, respectively, in terms of MOTA at 0.01bpp. We notice FVC poorly performs on MOT task, mainly due to its I frames are based on hand-crafted BPG codec [9] and P frames are based on neural codec. We conjecture the reason is that the tracking task heavily relies on consistent feature representation to lock on and keep track of the same object across different frames, and the I/P frame domain shift problem drastically confuses the tracktor. We also notice that PLVC achieves the second best result, and is even superior to the strong VVC codec, because the equipped GAN loss by PLVC right normalizes the object features into a more consistent space. Finally, we provide the RP curves on MOT task, as shown in the second row of Figure 2. It is observed that our method largely outperforms other ones.

Video Object Segmentation (VOS). We further benchmark the semantic coding performance of different methods on VOS task, which is more fine-grained than the MOT task and relies on pixel-level details to obtain accurate segmentation results. As shown in Table 3, our method achieves the best performance in terms of all metrics when compared to all baseline codecs. For example, our method outperforms VVC and PLVC by 6.73% and 12.75% at 0.01bpp level in terms of $\mathcal{J}\&\mathcal{F}$. We also illustrate the RP curves on VOS task. As shown in the third row of Figure 2, our framework consistently outperforms other methods.

	MOTA (%)↑	MOTP (%)↓	IDF1 (%)↑	FN↓
HEVC	61.30	19.88	64.32	17377
FVC	44.24	21.97	52.53	27508
PLVC	67.87	18.44	68.95	13299
VVC	64.86	19.49	68.99	15621
Ours	**70.84**	**17.79**	**71.89**	**11710**
Original	78.60	15.80	79.00	7000

Table 2: MOT performance comparison of different coding methods on MOT17 dataset at 0.01bpp. "Original" denotes the results with original videos, which is the performance upper bound.

	$\mathcal{J}\&\mathcal{F}$ (%)↑	$\mathcal{J}$ (%)↑	$\mathcal{F}$ (%)↑	$\mathcal{F}$-*Recal* (%)↑
HEVC	57.68	56.84	58.51	67.44
FVC	62.39	61.22	63.55	75.67
PLVC	61.45	60.02	62.87	74.07
VVC	67.47	65.59	69.36	80.92
Ours	**74.20**	**60.21**	**78.19**	**91.10**
Original	87.70	84.06	91.33	97.02

Table 3: VOS performance comparison of different coding methods on DAVIS2017 at 0.01bpp. "Original" denotes the results with original videos, which is the performance upper bound.

Video Quality. We further provide the rate-distortion (RD) curves of the proposed framework in Figure 3. Compared to perceptual quality-oriented video compression method PLVC [132] and traditional methods HEVC/VVC, our method achieves the best perceptual quality, which is indicated by the much lower LPIPS [141]. From the distortion-perception theory [12][13][52][129], the good perceptual coding capability is at the cost of high distortion, *i.e.*, the pixel-wise fidelity. Therefore, when evaluating with the PSNR metric, we fine-tune the decoder part of our framework by using the MSE loss as the distortion term. We refer to this model as "Ours*". It is worth noting that employing an alternative decoder for PSNR assessment does not compromise the practicality of our approach due to the bitstream is right the previously received one for AI tasks. Our method consistently outperforms HEVC/VVC codecs, as well as other recent PNSR-oriented learnable methods, *i.e.*, FVC [61], C2F [60] and DCVC [73]. For example, our improvement over VVC is 0.25db at 0.1bpp level.

To keep pace with the latest video compression methods, we further equip our method with the powerful al-

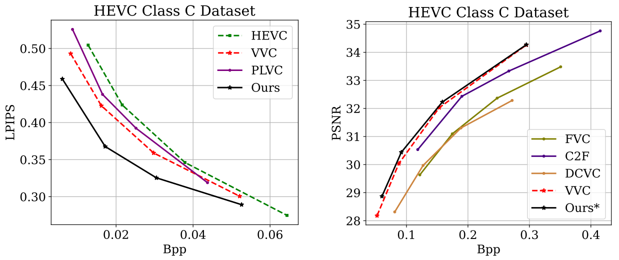

Figure 3: RD curves of different video coding methods.

beit slow VVC reference software VTM, end-to-end fine-tuning both our encoder and decoder networks with the RD loss. Our approach saves the bitcost of VTM17.0 by -11.2% on HEVC C in terms of BDBR [11], which is superior to the latest methods, *i.e.*, DCVC-DC [75] (-10.3%), DCVC-HEM [74] (+22.2%) and DCVC-TCM [97] (+66.3%).

4.4. Ablation Studies

Framework Study. We first train a variant model denoted by *SMC wo Comp*, in which the semantic compensation operation is removed. As shown in Figure 4 (*Left*), the performance is drastically decreased by 16% in terms of Top1 accuracy at 0.02bpp. Further, we replace the simple UNet-Style visual-semantic fusion network of *SMC wo Comp* with a compressed video enhancement method *BasicVSR++* [17]. The framework is degraded to a "lossy codec plus video enhancement" paradigm. The *BasicVSR++* model is dedicatedly trained for video compression artifact removal [18][130]. The improvement over the simple UNet is marginal and is still far behind our full semantic compensation framework by -14.23%@0.02bpp. These results strongly prove that the corrupted semantic information cannot be effectively fixed by video enhancement and should be particularly compensated.

We also fine-tune the official UCF101-TSM model using videos compressed by above variant methods. As shown in Table 4, the performance of vanilla VVC after fine-tuning (FT) nearly catches up with that of *BasicVSR++* (82.46%), suggesting that video enhancement procedure mainly narrows the domain gap between the evaluation and training video data input to TSM model, and bring no new information. In contrast, our full SMC framework further improves the accuracy from 86.46% (already higher than *BasicVSR++*) to 89.83% at 0.04bpp level. This proves that our method consistently benefits the downstream task by compensating new semantic information, no matter fine-tuning the downstream model or not.

Different Semantic Learning Objectives $\mathcal{L}_{Sem}$. In this section, we train several SMC models by equipping them with different learning objectives, *i.e.*, *MoCoV3* [27], *SegMap*, *HEDMap*, and *FeatureMatch*. The first one is purely self-supervised. The latter three ones regularize the compressed video to be similar to the original video in terms of DeepLabv3 [24] semantic segmentation map, HED [124] edge map, and VGG16 [99] feature map, respectively.

	VVC	*BasicVSR++*	Ours
Before FT(%)	76.97	82.46	**86.46**
After FT(%)	82.35(+5.38)	82.92(+0.46)	**89.83**(+3.37)

Table 4: Impact of fine-tuning (FT) downstream model to different coding methods on UCF101-TSM (Top1) at 0.04bpp level.

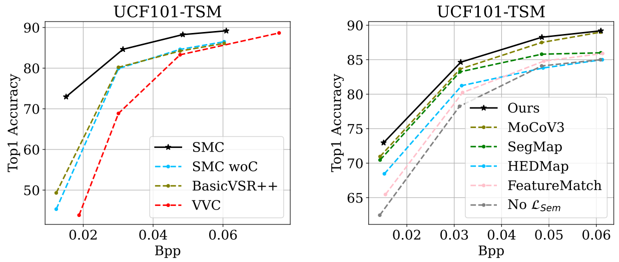

Figure 4: *Left:* Ablation studies on SMC framework. *SMC woC* denotes the semantic compensation procedure is removed. *BasicVSR++* denotes the compressed videos by VVC codec are enhanced by state-of-the-art BasicVSR++ method [17]. *Right:* Comparison of different semantic learning objectives.

UCF101-TSM@0.02bpp	MOT17-ByteTrack@0.01bpp
MoCoV3 → ***Ours***	*MoCoV3* → ***Ours***
Top1: 75.20% → **76.48%**	MOTA: 67.62% →**70.84%**(+3.22%)

Table 5: Comparison of different self-supervised learning methods on coarse/fine-grained tasks, *i.e.*, action recognition and MOT.

As shown in Figure 4 (*Right*), our method consistently outperforms the contrastive learning-based method *MoCoV3*. We also compare these two objectives on MOT task, the large performance gain (+3.22%) clearly proves our learned semantics is more generalizable. Then, we find that the hand-crafted *SegMap* loss achieves similar performance to *MoCoV3* in the lower bitrate ranges. This is consistent with the previous works [62][20], which relies on the hand-crafted design, *i.e.*, employing semantic segmentation map as the semantic intermediary, but still achieves promising results on downstream AI tasks. However, this paradigm is still far behind our learning objective, *e.g.*, about a 3% performance gap at 0.05bpp level. After replacing the segmentation map with the HED edge map, the performance is further dropped, because the edge map does not contain the category information of each object region and is of less semantics. The model trained with *FeatureMatch* shows the worst results, probably because feature values are denser than the discretized segmentation/edge maps, which increase the bitcost. When completely removing the $\mathcal{L}_{Sem}$ loss item, the performance is further degraded due to not particularly enforcing the video semantic completeness.

Effectiveness of NSS Strategy. Our method adapts the vanilla MAE to the semantic compression task by suppressing the non-semantic information within its token space. To study the necessity of this design, we first train a variant model *SMC woNSS* by removing the $\mathcal{L}_{NSS}$ item from the loss function. As shown in Table 6, the bitcost of *SMC woNSS* is 2.6× larger than the full SMC model, because the low-level pixel information is back-propagated to the coding system without any selection. Moreover, this

	SMC VQ	SMC woSem	SMC woNSS	SMC
Res (bpp)	0.0048	0.0042	0.0074	**0.0028**
Top1 (%)	71.22	71.38	69.57	**72.96**

Table 6: Comparison of different non-semantic suppression (NSS) strategies on UCF101-TSM. Res (bpp) denotes the bitcost of the residual semantic stream. The CRF value of VVC is set to 51.

non-semantic information may be noises to the downstream tasks, *i.e.*, the action recognition accuracy is dropped from 72.96% to 69.57%. Then, we train a variant model *SMC woSem* by using a plain learnable variable as the condition of the GMM distribution, instead of the semantic feature S. Both the compression efficiency and the recognition accuracy of *SMC woSem* are superior to *SMC woNSS*, but still inferior to our SMC model. This implies that explicitly regularizing the information entropy of the MAE token space is beneficial to a coding system, and using learned semantics as the guidance further improves this idea. Finally, we use the vector quantization (VQ) [114] codebook to discretize the token space of MAE, and the resulted *SMC VQ* model has similar performance to *SMC woSem*, indicating that the idea of information suppression is important for a compression system, instead of its concrete implementation.

Hyper-parameter Sensitivity. Setting the semantic loss item weight α to the values in range [1,10] gives the similar results. Setting the NSS loss item weight β to 0.1 achieve the best results, while smaller value 0.01 cannot effectively suppress the non-semantic information and larger value 1 makes the semantic learning item hard to optimize. Setting the GMM component number K to values in range [3,9] gives the similar results, while smaller K is slightly worse.

4.5. Model Analysis

Bit-allocation of Semantic and Visual Information. As shown in Figure 5, the semantic stream is always of high compression efficiency. Moreover, the bit allocation strategy is adaptive to the quality (CRF) of the lossy visual stream, although we do not introduce any explicit adaptive design and time-consuming online rate-distortion optimization (RDO) strategies like [81]. The proportion of semantic information has been decreased to about 1% when CRF is larger than 40 (the second column). This 1% bitcost boots Top1 accuracy by 7% on UCF101-TSM.

Visualization. As shown in Figure 6, the semantic feature extracted from the input frame is of high semantics and close to human perception. The activated regions are concentrated on the human body and the saliency objects, *i.e.*, vehicle wheels and drum. We further visualize the bitcost map of the residual part, where the redundant part is removed by subtracting the semantics in lossy video stream. The residual bitcost map is quite sparse, and further concentrated on the AI task-interested regions. We also compare the VVC codec and our method in terms of qualitative result. As shown in (e) and (f) of Figure 6, our method based on VVC (CRF=51) demonstrates much clearer object structures and sharper edges than vanilla VVC (CRF=47), while consuming few bits (0.08bpp *v.s.* 0.095bpp).

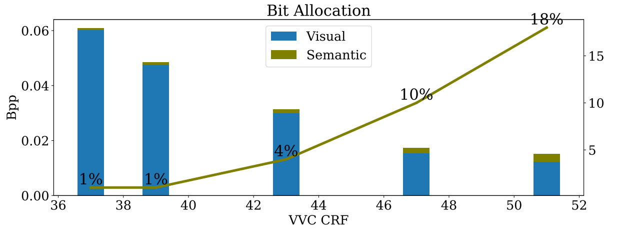

Figure 5: Bit consumption of visual and residual semantic information. The x-axis indicates the set CRF value of the VVC codec.

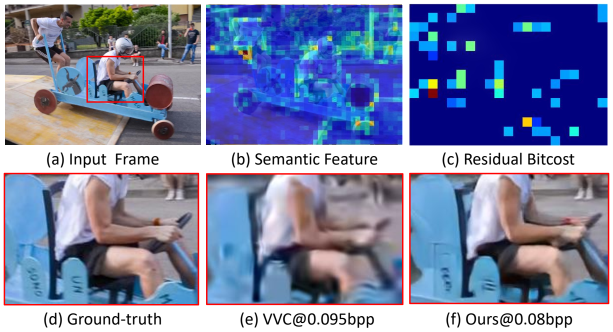

Figure 6: Visualization of the semantic feature and the bitcost map of the residual semantics. We also compare the visual quality of the zoomed-in region, which is denoted by the red box.

Model Complexity. The parameter numbers of Sem-Net, Sem-Net$_s$, F-Net and Density-Net are 12.2M, 3.1M, 8.4M and 1.3M, respectively. We report the per frame running time of a 1080p video on the machine with a Nvidia 2080Ti GPU. Our encoding time is 1413ms, of which 993ms is consumed by the VVCenc. Our decoding time is 446ms, of which 126ms is consumed by the VVCenc. Although our framework introduces extra decoding time for better supporting AI tasks, the frame decoded from visual stream can be directly analyzed and gives a fast response.

5. Conclusion and Limitation

In this paper, we have focused on a novel unsupervised video semantic compression problem. We have proposed a simple baseline framework SMC to better cope with this problem, which is equipped with a novel non-semantics suppressed MAE loss. We have also built a benchmark by evaluating several video codecs on three common video analysis tasks. Comprehensive experiments demonstrate our method achieves remarkable results. One limitation is the learned semantics still relies on the training dataset consisting of natural images, which may not perform well on professional field, such as medial image analysis.

Acknowledgement This work was supported by the National Science Foundation of China (62225112 and 61831015), National Key R&D Program of China (2021YFE0206700), STCSM under Grant 22DZ2229005 and 111 project BP0719010.

References

[1] Mmtracking: Openmmlab video perception toolbox and benchmark. https://github.com/open-mmlab/mmtracking, 2020. 5

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

[3] Mohammad Akbari, Jie Liang, and Jingning Han. Dsslic: Deep semantic segmentation-based layered image compression. In *ICASSP*, 2019. 2

[4] Yuanchao Bai, Xu Yang, Xianming Liu, Junjun Jiang, Yaowei Wang, Xiangyang Ji, and Wen Gao. Towards end-to-end image compression and analysis with transformers. In *AAAI*, 2022. 2

[5] Johannes Ballé, Valero Laparra, and Eero P Simoncelli. End-to-end optimized image compression. *arXiv*, 2016. 3

[6] Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston. Variational image compression with a scale hyperprior. *arXiv*, 2018. 3, 4

[7] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv*, 2021. 3

[8] Luca Baroffio, Matteo Cesana, Alessandro Redondi, Marco Tagliasacchi, and Stefano Tubaro. Hybrid coding of visual content and local image features. In *ICIP*, 2015. 1, 2

[9] F. Bellard. "the bpg image format," http://bellard.org/bpg/, last accessed on 09/20/2015. 6

[10] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? *arXiv*, 2021. 4, 5

[11] Gisle Bjontegaard. Calculation of average psnr differences between rd-curves. *VCEG-M33*, 2001. 6, 7

[12] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *CVPR*, 2018. 6

[13] Yochai Blau and Tomer Michaeli. Rethinking lossy compression: The rate-distortion-perception tradeoff. In *ICML*, 2019. 6

[14] 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. *TCSVT*, 2021. 1, 2

[15] Qi Cai, Zhifeng Chen, Dapeng Oliver Wu, Shan Liu, and Xiang Li. A novel video coding strategy in hevc for object detection. *TCSVT*, 2021. 2

[16] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017. 2, 4

[17] Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. In *CVPR*, 2022. 7

[18] Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. On the generalization of basicvsr++ to video deblurring and denoising. *arXiv*, 2022. 7

[19] Jianhui Chang, Qi Mao, Zhenghui Zhao, Shanshe Wang, Shiqi Wang, Hong Zhu, and Siwei Ma. Layered conceptual image compression via deep semantic synthesis. In *ICIP*, 2019. 2

[20] Jianhui Chang, Jian Zhang, Youmin Xu, Jiguo Li, Siwei Ma, and Wen Gao. Consistency-contrast learning for conceptual coding. In *ACMMM*, 2022. 2, 7

[21] Jianhui Chang, Zhenghui Zhao, Chuanmin Jia, Shiqi Wang, Lingbo Yang, Qi Mao, Jian Zhang, and Siwei Ma. Conceptual compression via deep structure and texture synthesis. *TIP*, 2022. 2

[22] Jianhui Chang, Zhenghui Zhao, Lingbo Yang, Chuanmin Jia, Jian Zhang, and Siwei Ma. Thousand to one: Semantic prior modeling for conceptual coding. In *ICME*, 2021. 2

[23] Jianshu Chao and Eckehard Steinbach. Keypoint encoding for improved feature extraction from compressed video at low bitrates. *TMM*, 2015. 1, 2

[24] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv*, 2017. 7

[25] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 2

[26] Xiaokang Chen, Mingyu Ding, Xiaodi Wang, Ying Xin, Shentong Mo, Yunhao Wang, Shumin Han, Ping Luo, Gang Zeng, and Jingdong Wang. Context autoencoder for self-supervised representation learning. *arXiv*, 2022. 3

[27] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv*, 2020. 2, 7

[28] Zhuo Chen, Kui Fan, Shiqi Wang, Lingyu Duan, Weisi Lin, and Alex Chichung Kot. Toward intelligent sensing: Intermediate deep feature compression. *TIP*, 2019. 1, 2

[29] Zhuo Chen, Kui Fan, Shiqi Wang, Ling-Yu Duan, Weisi Lin, and Alex Kot. Lossy intermediate deep learning feature compression and evaluation. In *ACM MM*, 2019. 1, 2

[30] Zhenghao Chen, Guo Lu, Zhihao Hu, Shan Liu, Wei Jiang, and Dong Xu. Lsvc: A learning-based stereo video compression framework. In *CVPR*, 2022. 2

[31] Ho Kei Cheng and Alexander G Schwing. Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. *arXiv*, 2022. 5

[32] Zhengxue Cheng, Heming Sun, Masaru Takeuchi, and Jiro Katto. Learned image compression with discretized gaussian mixture likelihoods and attention modules. In *CVPR*, 2020. 4

[33] Hyomin Choi and Ivan V Bajic. High efficiency compression for object detection. In *ICASSP*, 2018. 1, 2

[34] Hyomin Choi and Ivan V Bajić. Near-lossless deep feature compression for collaborative intelligence. In *MMSP*, 2018. 1, 2

[35] Hyomin Choi and Ivan V Bajić. Latent-space scalability for multi-task collaborative intelligence. In *ICIP*, 2021. 1, 2

[36] Hyomin Choi and Ivan V Bajic. Scalable image coding for humans and machines. *TIP*, 2022. 1, 2

[37] Jinyoung Choi and Bohyung Han. Task-aware quantization network for jpeg image compression. In *ECCV*, 2020. 2

[38] Yan Dai, Ziyu Hu, Shuqi Zhang, and Lianjun Liu. A survey of detection-based video multi-object tracking. *Displays*, 2022. 2

[39] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 2

[40] Xiaoyi Dong, Jianmin Bao, Ting Zhang, Dongdong Chen, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, and Nenghai Yu. Peco: Perceptual codebook for bert pretraining of vision transformers. *arXiv*, 2021. 3

[41] Xiaoyi Dong, Jianmin Bao, Ting Zhang, Dongdong Chen, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, and Nenghai Yu. Bootstrapped masked autoencoders for vision bert pretraining. In *ECCV*, 2022. 1, 2, 3

[42] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv*, 2020. 5

[43] Lingyu Duan, Jiaying Liu, Wenhan Yang, Tiejun Huang, and Wen Gao. Video coding for machines: A paradigm of collaborative compression and intelligent analytics. *TIP*, 2020. 2

[44] Ling-Yu Duan, Vijay Chandrasekhar, Jie Chen, Jie Lin, Zhe Wang, Tiejun Huang, Bernd Girod, and Wen Gao. Overview of the mpeg-cdvs standard. *TIP*, 2015. 1, 2

[45] Ling-Yu Duan, Feng Gao, Jie Chen, Jie Lin, and Tiejun Huang. Compact descriptors for mobile visual search and mpeg cdvs standardization. In *ISCAS*, 2013. 1, 2

[46] 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. *TMM*, 2018. 1, 2

[47] Shiyu Duan, Huaijin Chen, and Jinwei Gu. Jpd-se: High-level semantics for joint perception-distortion enhancement in image compression. *TIP*, 2022. 2

[48] Yann Dubois, Benjamin Bloem-Reddy, Karen Ullrich, and Chris J Maddison. Lossy compression for lossless prediction. *NIPS*, 2021. 2

[49] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, 2021. 3

[50] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *ICCV*, 2019. 2, 4

[51] 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. *arXiv*, 2022. 1, 2

[52] Dror Freirich, Tomer Michaeli, and Ron Meir. A theory of the distortion-perception tradeoff in wasserstein space. *NIPS*, 2021. 6

[53] Leonardo Galteri, Marco Bertini, Lorenzo Seidenari, and Alberto Del Bimbo. Video compression for object detection algorithms. In *ICPR*, 2018. 1, 2

[54] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 2020. 3

[55] Zongxi Han, Yutao Liu, and Rong Xie. A large-scale image database for benchmarking mobile camera quality and nr-iqa algorithms. *Displays*, 2023. 2

[56] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022. 1, 2, 3

[57] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 2

[58] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3

[59] 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 *ICME*, 2020. 2

[60] 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 *CVPR*, 2022. 6

[61] Zhihao Hu, Guo Lu, and Dong Xu. Fvc: A new framework towards deep video compression in feature space. In *CVPR*, 2021. 1, 2, 5, 6

[62] Danlan Huang, Xiaoming Tao, Feifei Gao, and Jianhua Lu. Deep learning-based image semantic coding for semantic communications. In *GLOBECOM*, 2021. 2, 7

[63] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, 2017. 3

[64] Zhimeng Huang, Chuanmin Jia, Shanshe Wang, and Siwei Ma. Visual analysis motivated rate-distortion model for image coding. In *ICME*, 2021. 2

[65] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. 3

[66] Rangachar Kasturi, Dmitry Goldgof, Padmanabhan Soundararajan, Vasant Manohar, John Garofolo, Rachel Bowers, Matthew Boonstra, Valentina Korzhova, and Jing Zhang. Framework for performance evaluation of face, text, and vehicle detection and tracking in video: Data, metrics, and protocol. *TPAMI*, 2008. 5

[67] Sayed Hossein Khatoonabadi and Ivan V Bajic. Video object tracking in the compressed domain using spatio-temporal markov random fields. *TIP*, 2012. 2

[68] Ahmad S Khattak, Nadeem Anjum, Nasrullah Khan, Muhammad R Mufti, and Naeem Ramzan. Amf-mspf: A retrospective analysis with online object tracking algorithms. *Displays*, 2022. 2

[69] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv*, 2014. 5

[70] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *ICCV*, 2011. 4

[71] Nam Le, Honglei Zhang, Francesco Cricri, Ramin Ghaznavi-Youvalari, and Esa Rahtu. Image coding for machines: an end-to-end learned approach. In *ICASSP*, 2021. 2

[72] 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 *CVPR*, 2022. 2

[73] Jiahao Li, Bin Li, and Yan Lu. Deep contextual video compression. *NIPS*, 2021. 2, 6

[74] Jiahao Li, Bin Li, and Yan Lu. Hybrid spatial-temporal entropy modelling for neural video compression. In *ACM MM*, 2022. 7

[75] Jiahao Li, Bin Li, and Yan Lu. Neural video compression with diverse contexts. In *CVPR*, 2023. 7

[76] Yingwei Li, Yi Li, and Nuno Vasconcelos. Resound: Towards action recognition without representation bias. In *ECCV*, 2018. 4

[77] Ji Lin, Chuang Gan, and Song Han. Tsm: temporal shift module for efficient video understanding. In *ICCV*, 2019. 4

[78] Hongyan Liu, Shuang Shi, Ruxue Bai, Yuchen Liu, Xinkang Lian, and Ting Shi. A brain-inspired computational model for extremely few reference image quality assessment. *Displays*, 2023. 2

[79] Kang Liu, Dong Liu, Li Li, Ning Yan, and Houqiang Li. Semantics-to-signal scalable image compression with learned revertible representations. *IJCV*, 2021. 1, 2

[80] Qiankun Liu, Bin Liu, Yue Wu, Weihai Li, and Nenghai Yu. Real-time online multi-object tracking in compressed domain. *arXiv*, 2022. 2

[81] Guo Lu, Chunlei Cai, Xiaoyun Zhang, Li Chen, Wanli Ouyang, Dong Xu, and Zhiyong Gao. Content adaptive and error propagation aware deep video compression. In *ECCV*, 2020. 8

[82] Guo Lu, Wanli Ouyang, Dong Xu, Xiaoyun Zhang, Chunlei Cai, and Zhiyong Gao. Dvc: An end-to-end deep video compression framework. In *CVPR*, 2019. 2

[83] Guo Lu, Xiaoyun Zhang, Wanli Ouyang, Li Chen, Zhiyong Gao, and Dong Xu. An end-to-end learning framework for video compression. *TPAMI*, 2020. 1, 2

[84] Guo Lu, Tianxiong Zhong, Jing Geng, Qiang Hu, and Dong Xu. Learning based multi-modality image and video compression. In *CVPR*, 2022. 2

[85] Di Ma, Fan Zhang, and David R Bull. Bvi-dvc: A training database for deep video compression. *TMM*, 2021. 2

[86] Mina Makar, Haricharan Lakshman, Vijay Chandrasekhar, and Bernd Girod. Gradient preserving quantization. In *ICIP*, 2012. 1, 2

[87] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *ICCV*, 2017. 3

[88] Fabian Mentzer, Eirikur Agustsson, Johannes Ballé, David Minnen, Nick Johnston, and George Toderici. Neural video compression using gans for detail synthesis and propagation. In *ECCV*, 2022. 2

[89] Anton Milan, Laura Leal-Taixé, Ian Reid, Stefan Roth, and Konrad Schindler. Mot16: A benchmark for multi-object tracking. *arXiv*, 2016. 4

[90] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv*, 2016. 3

[91] Tian Pan, Yibing Song, Tianyu Yang, Wenhao Jiang, and Wei Liu. Videomoco: Contrastive video representation learning with temporally adversarial examples. In *CVPR*, 2021. 2

[92] Athanasios Papoulis and S Unnikrishna Pillai. *Probability, random variables and stochastic processes*. 2002. 4

[93] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NIPS*, 2019. 5

[94] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv*, 2017. 4

[95] Douglas A Reynolds et al. Gaussian mixture models. *Encyclopedia of biometrics*, 741(659-663), 2009. 4

[96] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 3

[97] Xihua Sheng, Jiahao Li, Bin Li, Li Li, Dong Liu, and Yan Lu. Temporal context mining for learned video compression. *TMM*, 2022. 7

[98] 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 *CVPR*, 2019. 2

[99] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv*, 2014. 7

[100] Saurabh Singh, Sami Abu-El-Haija, Nick Johnston, Johannes Ballé, Abhinav Shrivastava, and George Toderici. End-to-end learning of compressible features. In *ICIP*, 2020. 1, 2

[101] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv*, 2012. 4

[102] Gary J Sullivan, Jens-Rainer Ohm, Woo-Jin Han, and Thomas Wiegand. Overview of the high efficiency video coding (hevc) standard. *TCSVT*, 2012. 1, 2, 4

[103] Simeng Sun, Tianyu He, and Zhibo Chen. Semantic structured image coding framework for multiple intelligent applications. *TCSVT*, 2020. 1

[104] Takehiro Tanaka, Alon Harell, and Ivan V Bajić. Does video compression impact tracking accuracy? *arXiv*, 2022. 1

[105] Yuan Tian, Zhaohui Che, Wenbo Bao, Guangtao Zhai, and Zhiyong Gao. Self-supervised motion representation via scattering local motion cues. In *ECCV*, 2020. 2

[106] Yuan Tian, Guo Lu, Xiongkuo Min, Zhaohui Che, Guangtao Zhai, Guodong Guo, and Zhiyong Gao. Self-conditioned probabilistic learning of video rescaling. In *ICCV*, 2021. 2

[107] Yuan Tian, Guo Lu, Yichao Yan, Guangtao Zhai, Li Chen, and Zhiyong Gao. A coding framework and benchmark towards compressed video understanding. *arXiv*, 2022. 2

[108] Yuan Tian, Xiongkuo Min, Guangtao Zhai, and Zhiyong Gao. Video-based early asd detection via temporal pyramid networks. In *ICME*, 2019. 2

[109] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? *NIPS*, 2020. 2

[110] Yuan Tian, Yichao Yan, Guangtao Zhai, Li Chen, and Zhiyong Gao. Clsa: a contrastive learning framework with selective aggregation for video rescaling. *TIP*, 2023. 2

[111] Yuan Tian, Yichao Yan, Guangtao Zhai, Guodong Guo, and Zhiyong Gao. Ean: event adaptive network for enhanced action recognition. *IJCV*, 2022. 2

[112] Suramya Tomar. Converting video formats with ffmpeg. *Linux Journal*, 2006. 5

[113] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *arXiv*, 2022. 1, 2, 5

[114] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *NIPS*, 2017. 8

[115] Anton Igorevich Veselov, Hu Chen, Francesco Romano, ZHAO Zhijie, and Marat Ravilevich Gilmutdinov. Hybrid video and feature coding and decoding, 2021. US Patent App. 17/197,500. 2

[116] Shiyao Wang, Hongchao Lu, and Zhidong Deng. Fast object detection in compressed video. In *ICCV*, 2019. 2

[117] Xinggang Wang, Zhaojin Huang, Bencheng Liao, Lichao Huang, Yongchao Gong, and Chang Huang. Real-time and accurate object detection in compressed video by long short-term feature aggregation. *CVIU*, 2021. 2

[118] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *ACSSC*, 2003. 2

[119] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *CVPR*, 2022. 3

[120] Adam Wieckowski, Jens Brandenburg, Tobias Hinz, Christian Bartnik, Valeri George, Gabriel Hege, Christian Helmrich, Anastasia Henkel, Christian Lehmann, Christian Stoffers, Ivan Zupancic, Benjamin Bross, and Detlev Marpe. Vvenc: An open and optimized vvc encoder implementation. In *ICMEW*. 5

[121] Thomas Wiegand, Gary J Sullivan, Gisle Bjontegaard, and Ajay Luthra. Overview of the h. 264/avc video coding standard. *TCSVT*, 2003. 2

[122] Chao-Yuan Wu, Nayan Singhal, and Philipp Krahenbuhl. Video compression through image interpolation. In *ECCV*, 2018. 4

[123] Chao-Yuan Wu, Manzil Zaheer, Hexiang Hu, R Manmatha, Alexander J Smola, and Philipp Krähenbühl. Compressed video action recognition. In *CVPR*, 2018. 2

[124] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *ICCV*, 2015. 7

[125] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *CVPR*, 2022. 3

[126] Ning Yan, Changsheng Gao, Dong Liu, Houqiang Li, Li Li, and Feng Wu. Sssic: Semantics-to-signal scalable image coding with learned structural representations. *TIP*, 2021. 1, 2

[127] Zhicong Yan, Gaolei Li, Yuan Tian, Jun Wu, Shenghong Li, Mingzhe Chen, and H Vincent Poor. Dehib: Deep hidden backdoor attack on semi-supervised learning via adversarial perturbation. In *AAAI*, 2021. 2

[128] Zhicong Yan, Shenghong Li, Ruijie Zhao, Yuan Tian, and Yuanyuan Zhao. Dhbe: Data-free holistic backdoor erasing in deep neural networks via restricted adversarial distillation. In *AsiaCCS*, 2023. 2

[129] Zeyu Yan, Fei Wen, Rendong Ying, Chao Ma, and Peilin Liu. On perceptual lossy compression: The cost of perceptual reconstruction and an optimal training framework. In *ICML*, 2021. 6

[130] Ren Yang. Ntire 2021 challenge on quality enhancement of compressed video: Methods and results. In *CVPRW*, 2021. 7

[131] Ren Yang, Fabian Mentzer, Luc Van Gool, and Radu Timofte. Learning for video compression with hierarchical quality and recurrent enhancement. In *CVPR*, 2020. 3

[132] Ren Yang, Luc Van Gool, and Radu Timofte. Perceptual learned video compression with recurrent conditional gan. *arXiv*, 2021. 2, 3, 5, 6

[133] Shuai Yang, Yueyu Hu, Wenhan Yang, Ling-Yu Duan, and Jiaying Liu. Towards coding for human and machine vision: Scalable face image coding. *TMM*, 2021. 1, 2

[134] Zhaohui Yang, Yunhe Wang, Chang Xu, Peng Du, Chao Xu, Chunjing Xu, and Qi Tian. Discernible image compression. In *ACMMM*, 2020. 2

[135] Chenyu Yi, Siyuan Yang, Haoliang Li, Yap-peng Tan, and Alex Kot. Benchmarking the robustness of spatial-temporal models against corruptions. *NIPS*, 2021. 1

[136] Fuwang Yi, Mianyi Chen, Wei Sun, Xiongkuo Min, Yuan Tian, and Guangtao Zhai. Attention based network for no-reference ugc video quality assessment. In *ICIP*, 2021. 2

[137] Bowen Zhang, Limin Wang, Zhe Wang, Yu Qiao, and Hanli Wang. Real-time action recognition with enhanced motion vector cnns. In *CVPR*, 2016. 2

[138] Fan Zhang and David R Bull. A parametric framework for video compression using region-based texture models. *JSTSP*, 2011. 2

[139] Fan Zhang, Chen Feng, and David R Bull. Enhancing vvc through cnn-based post-processing. In *ICME*, 2020. 2

[140] Qi Zhang, Shanshe Wang, Xinfeng Zhang, Siwei Ma, and Wen Gao. Just recognizable distortion for machine vision oriented image and video coding. *IJCV*, 2021. 2

[141] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 3, 6

[142] Xiang Zhang, Siwei Ma, Shiqi Wang, Xinfeng Zhang, Huifang Sun, and Wen Gao. A joint compression scheme of video feature descriptors and visual content. *TIP*, 2016. 1, 2

[143] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Byte-

track: Multi-object tracking by associating every detection box. *arXiv*, 2021. 5

[144] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *arXiv*, 2021. 2, 3

[145] Wei Zhou, Zhibo Chen, and Weiping Li. Dual-stream interactive networks for no-reference stereoscopic image quality assessment. *TIP*, 2019. 2

[146] Wei Zhou, Likun Shi, Zhibo Chen, and Jinglin Zhang. Tensor oriented no-reference light field image quality assessment. *TIP*, 2020. 2

[147] Wei Zhou, Jiahua Xu, Qiuping Jiang, and Zhibo Chen. No-reference quality assessment for 360-degree images by analysis of multifrequency information and local-global naturalness. *TCSVT*, 2021. 2

[148] Chen Zhu, Guo Lu, Rong Xie, and Li Song. Perceptual video coding based on semantic-guided texture detection and synthesis. In *PCS*, 2022. 2