Deep Learning in Latent Space for Video Prediction and Compression

Bowen Liu Yu Chen Shiyu Liu Hun-Seok Kim
University of Michigan, Ann Arbor
{bowenliu, unchenyu, shiyuliu, hunseok}@umich.edu

Abstract

Learning-based video compression has achieved substantial progress during recent years. The most influential approaches adopt deep neural networks (DNNs) to remove spatial and temporal redundancies by finding the appropriate lower-dimensional representations of frames in the video. We propose a novel DNN based framework that predicts and compresses video sequences in the latent vector space. The proposed method first learns the efficient lower-dimensional latent space representation of each video frame and then performs inter-frame prediction in that latent domain. The proposed latent domain compression of individual frames is obtained by a deep autoencoder trained with a generative adversarial network (GAN). To exploit the temporal correlation within the video frame sequence, we employ a convolutional long short-term memory (ConvLSTM) network to predict the latent vector representation of the future frame. We demonstrate our method with two applications; video compression and abnormal event detection that share the identical latent frame prediction network. The proposed method exhibits superior or competitive performance compared to the state-of-the-art algorithms specifically designed for either video compression or anomaly detection. [1]

1. Introduction

Video data transmission occupies the majority of the internet data traffic nowadays. With the trend of extensive mobile devices usage worldwide, video data streaming is extensively used for productivity tools and entertainment platforms that assist people's work and life in various aspects. On top of the ubiquitous video engagement, superior video quality standards such as 4k UHD, and VR 360 became more widely available, which makes high performance video compression even more critical. Traditional video coding standards such as MPEG, AVC/H.264 [49], HEVC/H.265 [43], and VP9 [38] have achieved impressive performance on video compression tasks. However, as their primary applications are human perception driven, those hand-crafted codecs are likely suboptimal for machine-related tasks such as deep learning based video analytic.

Figure 1. Reconstructed frame with the conventional codecs (H.264, H.265) and our approach. Information and details are well preserved in the frame generated from a purely prediction-based latent representation (top right). Compared with H.264, our result yields less block artifacts and preserves finer details. Our method achieves a higher compression ratio than H.265 with similar quality.

During recent years, a growing trend of employing deep neural networks (DNNs) for image compression tasks has been witnessed. Prior works [46, 7, 36] have provided theoretical basis for application of deep autoencoders (AEs) on image codecs that attempt to optimize the rate-distortion trade-off, and they have showed the feasibility of latent representation as a format of compressed signal. While image compression reduces the redundancy only in spatial domain, video compression exploits the temporal correlation among consecutive frames as well. Using learned video prediction to substitute traditional block-based motion prediction/estimation methods has become a critical part of deep learning based video compression. Related recent works [14, 25, 27] address the uncertainties of real-world videos with stochastic video prediction networks using autoencoders and/or generative adversarial network (GAN) structures in recurrent settings. Learned video compression is a relatively recent topic. Early works [11, 50] either directly interpolate the key-frames or emulate the functional units in hand-crafted codecs with neural networks. Later proposed deep neural video codecs [31, 28, 19, 15, 41, 3, 51, 17, 30, 21] mainly target on learning data-driven algorithms that take advantage of the end-to-end trainability of DNNs. Most of them [28, 19, 15, 41, 3, 51, 17] adopt autoencoder style structures that encode frame and residual representations in

[1] Code available at: *https://github.com/BowenL0218/Video-compression*

latent space.

In this paper, we present a novel end-to-end deep learning video codec that benefits from video prediction in latent space. Our proposed method obtains the compressed frames in latent space by searching for the optimal latent representations [26], and then it learns temporal correlation within the latent space sequential data under a recurrent network setting. As oppose to previous approaches, the training and inference processes of our proposed prediction network are entirely performed in the latent domain. Our video coding method share the same predictor between the sender and receiver, and only transmit (store) the quantized and entropy coded prediction error (residual). The residual corrected latent frames are fed back to the prediction network for progressive estimation on consecutive latent representations of the data sequence.

Video compression evaluation results validate that this technique achieves superior performance compared to the state-of-the-art video codes. With the proposed prediction method, we also demonstrate its application on abnormal event detection, which is triggered when the prediction error exceeds a predefined threshold that represents a normal event. Anomaly detection evaluation results confirm superiority/competitiveness of the proposed method compared to recent algorithms specifically designed for that task.

Our main contributions are summarized as follows:

• **An end-to-end learned video compression codec based on GAN-prior generative image compression:** We propose a novel approach for time-series data compression by adopting a GAN-based autoencoder architecture in company with trainable quantization and entropy coding. Our codec provides a wider range of rate-distortion trade-off than what other recent (learning-based) codecs can offer.

• **Learning based video prediction in latent domain:** We use an convolutional long short-term memory (ConvLSTM) network to predict a compact latent representation of the next frame substituting for motion compensation in conventional codecs. This approach only stores the differences between the predicted and actual representation in low dimensional latent space, resulting in entropy reduction of the residuals. The predictor is adversarially trained against a discriminator which significantly enhances the quality of prediction to bring down the entropy (*i.e.*, density and magnitude of non-zero elements) of residuals.

• **Demonstration of a perceptual task in latent domain:** We investigate the feasibility and effectiveness of our video prediction algorithm by performing perceptual tasks in latent space. We showcase anomaly detection using the same ConvLSTM predictor designed for video compression. With unlabeled actions in the event detection dataset, our framework performs unsupervised learning on the video content and demonstrates reliable anomaly detection capability.

2. Related works

Learning-based image compression. There has been extensive study on applying DNNs to image compression tasks. Most approaches typically seek compression gain from translating images to lower-dimensional representations through either recurrent neural networks (RNNs) [45, 6, 23, 46] or autoencoder style networks [7, 44, 8, 34, 36]. Recent approaches use GAN-based structures for image compression [42, 4, 26] aiming to enhance subjective quality of image reconstructions from deep encoder-decoder pairs and take advantage of the qualification feedback provided by a discriminator. These approaches often target on optimizing distortion indicators such as mean squared error (MSE), PSNR, and MS-SSIM between the raw and reconstructed image, or the hybrid objective function including the perceptual loss. Our work adopts a GAN based structure to search for the optimal latent vectors that minimize distortion via back-propagation through a pre-trained generator (decoder).
Learning-based video compression. Video coding benefits from exploiting the temporal correlation between subsequent frames. Similar to the conventional codecs, learned video compression leverages the temporal correlation through inter-frame prediction. Chen *et al.* [11] first predict a frame then encode the residual (error between the prediction and actual) with a CNN. This approach shares great similarities with the block-based codecs. Arguably, however, DNNs are less efficient to learn from small image blocks. To overcome that issue, Wu *et al.* [50] propose a codec that captures temporal redundancy through hierarchical interpolation between key frames. The method uses a non-DNN based optical flow to generate motion information, and it is not jointly optimized with the rest of the model. Lu *et al.* [31] construct a DNN-based video compression pipeline close to the conventional codecs and optimize compression rate in conjunction with distortion. Lombardo *et al.* [28] present a learning-based video codec that performs end-to-end optimization on rate-distortion trade-off, quanitzation, and entropy coding. The framework is built with sequential variational autoencoder (VAE) where the encoded global state based prediction is used to tackle the temporal redundancy. Similarly, Rippel *et al.* [41] propose to represent all prior memory as a generic and learnable state that will continuously be updated during its propagation. The flow-residual information between two consecutive frames is generated from the state representation. Habibian *et al.* [19] use a 3D spatiotemporal autoencoder network that temporally decorrelates the latent vectors. Based on the encoder-decoder pair proposed in [7] for image compression, Djelouah *et al.* [15] encode displacement and blending coefficients into latent space representations. To address the failure cases typically observed in the flow-residual paradigm, Agustsson *et al.* [3] propose a scale-space flow that trilinearly warps the frame stack constructed by the previous frame as well as its variations obtained by applying

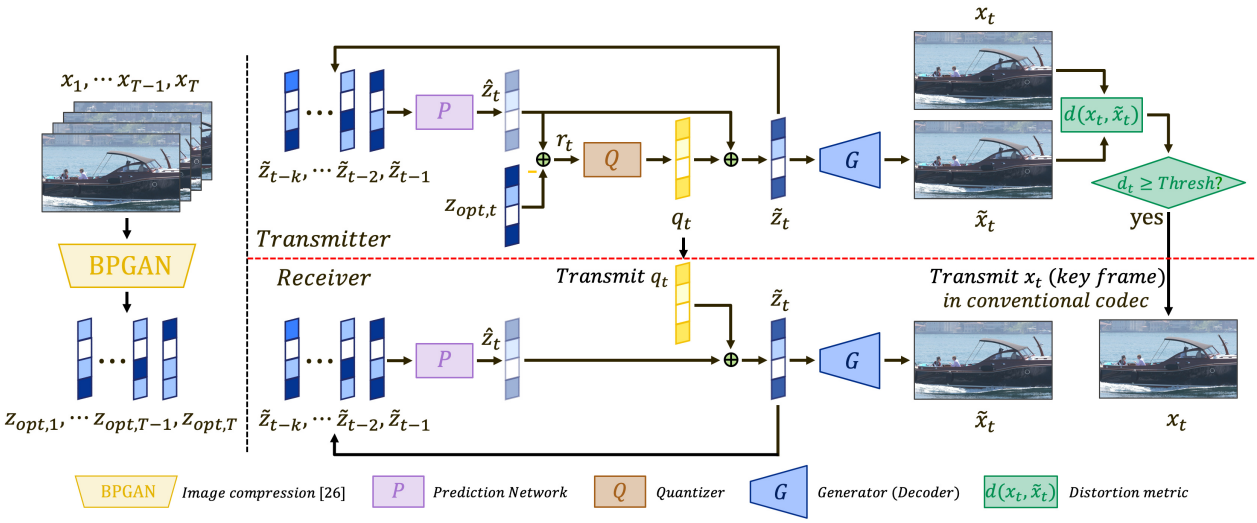

Figure 2. The operational flow of our proposed codec. Each frame of the video sequence is first individually compressed to a latent representation by the method in [26]. The predicted next-frame latent (output of the prediction network) is conditioned on the former reconstructed latent representations. The data we store and transmit is the quantized and entropy coded residual q_t. The transmitter and the receiver share the same prediction network, which produces identical reconstructed latents $\tilde{z}_t$ on both sides. $\tilde{z}_t$ is fed to the prediction network to estimate future frames, and to the generator to decode the video frame $\tilde{x}_t$. To preserve the quality of the reconstructed data sequences, we monitor the distortion between the original and reconstructed frames on the transmitter side, and directly send the encoded original key-frame to the receiver only if the distortion is above a certain threshold.

different level of Gaussian blurring. Following the hierarchical prediction approach, Yang *et al.* [51] design a framework that encodes video frames with different quality levels, and refines the coarsely predicted frames by leveraging the temporal correlation contained in the high quality frames. In our work, unlike the prior works mentioned above, the spatial redundancy is primarily exploited by finding the optimal low dimensional latent vectors to represent each video frame. Then we perform temporal predictions on successive frames in latent space. The residuals between the directly compressed frames (*i.e.*, latent vectors) and the predicted ones are quantized, entropy coded, and transmitted to the receiver.

Video prediction and motion compensation. The study on deep neural video prediction has led to a number of design choices. Early works usually devote to predicting small frame patches. To reduce the blurry reconstruction effect, Mathieu *et al.* [33] train a multi-scale network in an adversarial setting. Whereas Finn *et al.* [16] present a LSTM based network to learn the motion dynamics and to construct motion information with the content mask to form a predicted frame. Other approaches such as [5, 14, 25] propose variational methods to address the embedded stochasticity in real-world videos. Motion compensated prediction is an essential sub-task of video compression. Chen *et al.* [11] present a DNN based implementation that resembles block motion estimation in traditional codecs while others [31, 19] incorporate an optical flow encoder network into the compression system. Unlike prior approaches, our framework employs ConvLSTM based frame prediction in latent space for motion compensation. With a well-learned prediction network, we demonstrate that very sparse residuals can be obtained in latent space to produce extremely compressed video sequences.

Anomaly in the scene detection. Anomaly detection can be treated as an application of video prediction. A network structure in [13] includes cascaded convolutional LSTM networks in the autoencoder to learn the spatio–temporal features of the video frames. Liu *et. al* [27] are the first to introduce a video prediction framework adversarially trained under a temporal constraint for anomaly detection. Park *et. al* [39] propose to further enhance the performance by adopting a memory module to record representative normal patterns. Different from these works, our approach targets on predicting the next-frame in latent domain. In this work we demonstrate that the video representation learnt from latent space temporal redundancy can be adopted to perform reliable abnormal event detection.

3. Method specification

3.1. Video compression framework

Figure 2 depicts the proposed video compression framework. We first encode each frame to an optimal latent representation using the technique in [26]. This image compression technique searches for an optimal latent representation through the frame-by-frame back-propagation using a pre-trained generator network. We trained the generator (which serves as the decoder in the proposed video coding frame-

work) such that it can reconstruct a close-to-original frame from a latent representation. Once the optimal latent representation is produced for each frame, our end-to-end video compression framework learns temporal correlation among the latent space representations of consecutive video frames. To achieve this goal, we predict the next-frame's latent representation based on the sequence of latents of previous frames using a ConvLSTM. The prediction network takes the optimal latent vectors of each frame as the input and it is trained to predict the latent vector for the next frame as close as possible to the actual one. The element-wise difference between the predicted and actual latent is stored as the residual. Given a successfully trained latent space predictor, the residual is sparse with low entropy. Hence we attain the inter-frame compression gain from prediction on top of the intra-frame compression of compact latent representations.

To further reduce the video code size, we encode the residual with quantization and entropy coding. A desired compression rate is controlled by the size of latent dimension in the image compression stage as well as the number of quantization levels used in residual encoding. The quantized and entropy-coded residuals are sent from the transmitter to the receiver as the compressed representation of the video.

The reconstructed latent is obtained by adding the compressed residual to the predicted latent. The transmitter and the receiver share the same prediction network, which produces the identical reconstructed latent for the next frame using the previously reconstructed latent frames. At the beginning stage of our proposed video compression flow, the predictor on both sides is initialized with the latent vectors of several *initial frames* ($z_{opt,1:k}$, with $k = 6$ in our experiment) that are generated without prediction. This ensures the prediction for successive latents on both sides starts with the same recurrent state. Using the same generator (decoder) adopted in the image compression stage, the reconstructed latent ($\tilde{z}_t$) is translated to the spatial domain video frame, x_t.

We formulate the image compression problem as a joint model of a raw image x and its discrete latent representation z with θ representing model parameters:

$$p_\theta(x, z) = p_\theta(z) \cdot p_\theta(x|z) \tag{1}$$

In the above formula, $p_\theta(x|z)$ is the prior model and $p_\theta(z)$ is the likelihood. Under the scheme of video compression, the proposed ConvLSTM prediction network exploits temporal correlation, resulting in a likelihood model given the former latent representations in the sequence. Therefore, the prior model and the likelihood expression can be redefined as

$$p_\theta(x_{1:T}, z_{1:T}) = \prod_{t=1}^{T} p_\theta(z_t|z_{<t}) \cdot p_\theta(x_t|z_t) \tag{2}$$

where t is the time index for a frame. In the following sections, we separately address the main functional units of the proposed framework.

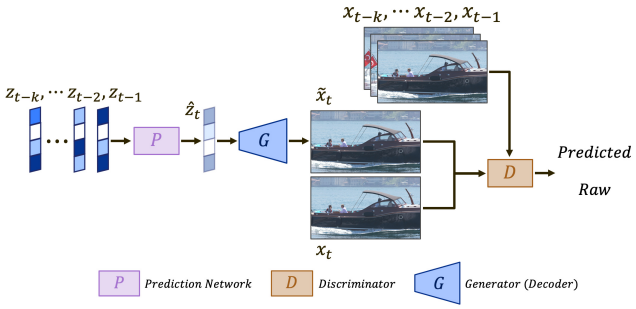

Figure 3. The latent space video prediction network estimates the next-frame latent with the understanding of latent space temporal correlation. A discriminator conditioned on the preceding frames provides feedback to the predictor so that it learns the mapping from the latent domain to the image space as well as the correlation of the previous and current frames. Given the sequence $x_{t-k}, ..., x_{t-1}$, the actual x_t and predicted $\tilde{x}_t$ are labeled as real and fake samples for the discriminator. As such, the predictor adversarially learns the temporal correlation.

3.2. Video prediction

The adopted image compression method [26] provides an optimal lower-dimensional representation z_{opt} in latent space for each image by minimizing a distortion function. As such, the image compression model learns a lossy transformation from spatial domain to latent space. Temporal correlation between subsequent frames in latent domain is exploited with an ConvLSTM based predictive model. An accurate inter-frame prediction model is a critical component in time-series data compression to capture the temporal correlation in the frame sequence and thereby to achieve small cross-entropy between the original and compressed data. A well-trained predictor learns the capability to predict the normal inter-frame content of the video such as the movement of an object and the translation of the camera. This characteristic of the predictor allows abnormal event detection as a byproduct of video compression as discussed in Section 4.5.

Similar to existing video codecs, our approach only encodes and transmits the residual between the predicted latent vector and the optimal latent z_{opt} obtained from the image compression process. Next frame prediction is hinging on the conditional prior model learned by the prediction network, whose cells implement the memory of previous data distribution. The generic prior model is defined as

$$p_\theta(z_T|z_{<T}) = \prod_{t=2}^{T} \frac{p_\theta(z_{1:t})}{p_\theta(z_{1:t-1})}. \tag{3}$$

Given the complexity of stochastic data distribution in videos and the possible rapid transition between frames, training a good prior model is a main challenge to obtain satisfying compression performance. We propose a GAN-based adver-

sarially trained ConvLSTM network to make predictions $\hat{z}$ on z_{opt} for video frame reconstruction.

As opposed to previously proposed methods, our prediction model $P(\cdot)$ produces estimated latent vectors instead of frames. This method requires significant attention to define a proper reconstruction objective function since element-wise error in latent space is often insufficient to measure spatial domain image reconstruction fidelity. The proposed prediction network is trained under an adversarial setup to exploit the complex similarity metric implicitly learned by a discriminator [24], which plays a critical role under the latent space prediction regime. Involving the discriminator objective [37] in the cost function improves the LSTM cells to establish more effective memory in terms of learning the temporal correlation. The loss function in our framework is expressed as

$$\begin{aligned}\mathcal{L} = \lambda \cdot \Big\{ & \mathbb{E}_{z\sim p_z}\big[\log\left(1 - D(G(P(z_{<t}|x_{<t})))\right)\big] \\ & + \mathbb{E}_{z\sim p_{opt}}\big[\log\left(D(G(z_{opt}|x_{<t}))\right)\big]\Big\} \\ & + (1-\lambda)\cdot \mathbb{E}_{p(z<t)}\big[\log p(z_t|z_{<t})\big],\end{aligned} \tag{4}$$

where $G(\cdot)$ is the generator network that reconstructs a frame x from a latent z and $D(\cdot)$ is the discriminator network that judges whether a frame belongs to a valid frame sequence as shown in Figure 3. The first term of the above loss function refers to the cross entropy of the discriminator cost function and the second term indicates the prediction error. Crucially, the image and video compression frameworks share the same generator (decoder) to reconstruct a frame x from a latent z. The generator parameters are inherited from image compression thus fixed when training the prediction network. As shown in Figure 2, the reconstructed latent vectors, $\tilde{z} = \hat{z} + Q(r)$, inevitably lose some information from the optimal ones, $z_{opt} = \hat{z} + r$, due to discretization $Q(r)$. Note that the quantization distortion metric is not presented in the prediction network loss, thus a well-trained generator $G(z)$ is crucial to minimize the loss between the target frame x_t and $\tilde{x}_t$ synthesized/generated from $\tilde{z}_t$.

3.3. Quantization and Entropy Coding

Now we describe the quantization and entropy coding of the residual r_t from latent prediction incorporated in the proposed video compression framework.

Quantization: We apply ADMM [40] quantization with a goal to find discretized vectors with minimal degradation of quality compared to the original frames. A generic quantization problem can be described as follows:

$$\min_r f(r) \quad \text{subject to } r \in S, \tag{5}$$

where S is a quantized set and f is a loss function. In the context of residual quantization, the loss function (6) is defined as

$$f(r) = d_l(z_{opt}, (\hat{z}+r)) + d_s(x, G(\hat{z}+r)) \quad \text{subject to } r \in S. \tag{6}$$

The optimization problem above is given by a combination of a) the latent space distortion d_l given the optimal z_{opt}, and b) the distortion $d_s(\cdot)$ in spatial domain reconstruction. It is non-convex and not solvable with stochastic gradient descent (SGD) due to the quatization constraint. Moreover, direct quantization is likely to cause gradient vanishing. Ren *et al.* [40] address these issues with ADMM using an indicator function and combining it with a differentiable loss function. During iterations, ADMM method projects all elements of residual latent vectors to different quantized levels and minimizes the loss function in parallel. This guarantees that all elements are quantized in the process to find the optimal level.

In ADMM quantization, the problem (5) is redirected to optimizing the cost function $\min_r f(r) + g(u)$ subject to $r = u$ by introducing u as an auxiliary variable. The indicator function $g(u)$ is 0 if $u \in S$ or ∞ otherwise.

$$\min_r f(r) + g(u) \quad \text{subject to } r = u \tag{7}$$

$$g(u) = \begin{cases} 0 & \text{if } u \in S \\ +\infty & \text{otherwise} \end{cases} \tag{8}$$

Then, the augment Largrangian method decomposes the dual variable optimization problem into two partial updating tasks performed iteratively and separately as described in (9), (10), and (11). With the addition of a convex and differentiable regularization term, the optimal solution is iteratively approximated through SGD with Adam optimizer.

$$r_{k+1} = \arg\min_r f(r) + \frac{\mu}{2}\cdot \|r - u_k + \eta_k\|_2^2 \tag{9}$$

$$u_{k+1} = \arg\min_u g(u) + \frac{\mu}{2}\cdot \|r_{k+1} - u + \eta_k\|_2^2 \tag{10}$$

$$\eta_{k+1} = \eta_k + r_{k+1} - u_{k+1} \tag{11}$$

Eq. (10) is solved by the Euclidean projection of $r_{k+1} + \eta_k$ onto the quantized set S, which is formulated as $u_{k+1} := \Pi_S\left(r_{k+1} + \eta_k\right)$ where Π_S is the projection function.

In our proposed method, we adopt an adaptive non-uniform quantization scheme where the quantization set is determined specifically for each video clip and transmitted together with the compressed data. Non-uniform quantization is implemented by selecting a subset of uniformly quantized levels and we only store/transmit the entropy coded indices of the selected quantization levels instead of the original value.

Entropy coding: A well-trained prediction model produces very sparse (with few non-zero elements) residual latent representations after quantization. The quantized residual vector is reshaped and stored in the compressed sparse row format

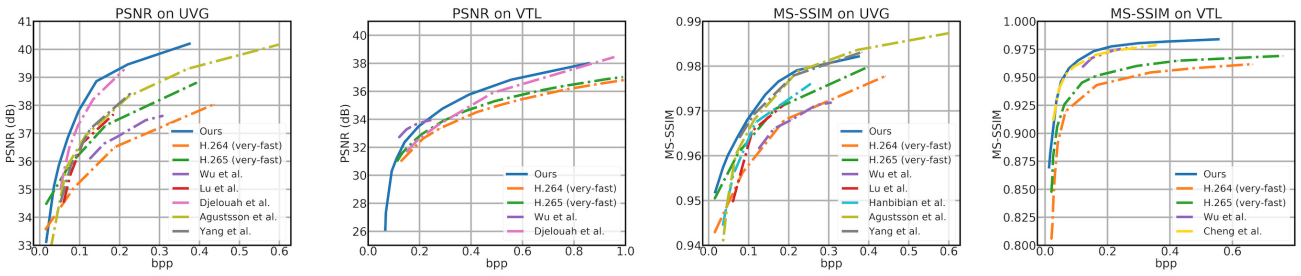

Figure 4. Comparison of our proposed approach with H.264, H.265, and learning-based codecs [50, 31, 15, 51, 3] for PSNR, and [50, 31, 19, 12, 51, 3] for MS-SSIM on UVG and VTL datasets. Our video codec is not optimized specifically for PSNR or MS-SSIM.

and finally entropy coded with Adaptive Arithmetic Coding [22]. After this final step, our codec achieves extremely high compression ratios with superior/competitive reconstruction quality compared to conventinal codecs.

3.4. Rate-distortion control

In order to allow a wide range of rate-distortion trade-off, the compression rate (or bit-per-pixel, bpp) of our method is controlled by changing the number of elements in the latent vector and the number of quantization levels for the residuals. The transmitter in our approach continuously monitors the quality of the reconstructed (decompressed) frame using a distortion metric $d(x_t, \tilde{x}_t)$ shown in Figure 2 to adjust the compression method on-the-fly. Note that the prediction based residual encoding can occasionally fail when the scene changes abruptly. In rare occasions, the generator (decoder) may yields inferior reconstructed frame $\tilde{x}_t$ due to limitations of the trained generator model. The frames that cause these issues are defined as *key-frames* and we encode key-frames using a conventional image codec BPG [9]. These key-frames are equivalent to intra-coded frames in conventional video codecs. This adaptive encoding prevents catastrophic failures in the proposed method that is designed to cover a wide range of rate-distortion trade-off space.

4. Experiments

4.1. Experiment setup

In the previous section, we introduced the loss function (eq. 4) that combines the loss of the latent reconstruction and the discriminator loss to enhance the quality of the reconstructed image from our prediction network. We empirically choose $\lambda = 0.1$ during the training to balance the first and second term. Adding the discriminator loss term does not necessarily improve the PSNR/SSIM metric but it does enhance the subjective quality of video/image. The number of iterations typically needed for ADMM quantization in our task is 50. The structure of DNNs (including the number of layers and kernel sizes, etc.) used in the experiment is specified in the supplementary material.

4.2. Video compression: datasets, metrics, and analysis

Datasets: Our framework is trained with the Kinetics dataset [10] and the UGC dataset [48]. In the Kinetics dataset, we use roughly 98,000 videos each lasting for around 10 seconds with resolution higher than $720p$. The UGC dataset has a rich collection of contents such as lecture, animation, and music videos with more than 1500 clips for an average length of 20 seconds. Training and evaluation datasets are mutually exclusive. We evaluate our approach using Video Trace Library (VTL) [1] and Ultra Video Group (UVG) [35] datasets. The VTL dataset contains 20 videos with around 40,000 frames of resolution 352×288. The UVG dataset has 16 videos, and we test on the original 8 videos with overall 3,900 frames of resolution 1920×1080 to compare with other existing methods.

Metrics: We quantitatively compare our experimental results with the most prevailing hand-crafted video codecs as well as learning-based codecs recently proposed. The quality distortion is measured by PSNR and MS-SSIM of decompressed frames. For the conventional codecs, H.264 and H.265, we use the *ffmpeg* very-fast mode with a GOP of 10/12 for the VTL/UVG dataset.

Analysis: The experimental results on VTL and UVG dataset in Figure 4 show that our work outperforms AVC/H.264, HEVC/H.265, and the state-of-the-art DNN based codecs for the most of tested bits-per-pixel (bpp) rates in terms of PSNR and MS-SSIM. Figure 1 is the visualization of a frame from the UVG dataset showing that our approach gives equally if not more visually appealing reconstructed frames under a lower/same bpp compared with other codecs. Additional visualization results from our method are available in the supplementary material. We observed from the output that incorporating discriminator loss in training can provide more complex details and realistic quality. Although this subjective quality gain is not always captured by the commonly adopted pixel-wise distortion metric such as PSNR and MS-SSIM, our method still achieves better results in the rate-distortion tradeoff measured by PSNR and MS-SSIM compared to other video codecs. Note that our scheme is flexible to adopt a different set of $z_{opt,t}$ that

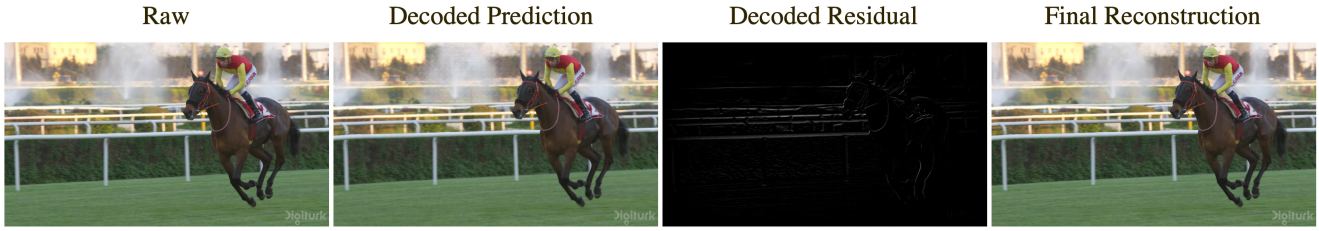

Figure 5. Visualization of our model results. The prediction network predicts for the next-frame latent according from several preceding latents (second left). The residual (second right) in the latent space is added to the predicted latent for final reconstruction via the generator.

minimizes an application-specific target metric (instead of generic MSE) such as a CNN feature loss or discriminator loss if the decompressed video is intended for machine learning based applications. We observed our PSNR in a very low bpp regime can be lower than that of some other codecs (Figure 4 second left). It is mainly because of two factors: 1) there is resolution mismatch between training and evaluation datasets, and 2) our target z minimizes a combined loss as shown in eq.(4) for superior subjective quality on reconstructed frames although it may result in slightly degraded PSNR.

4.3. Latent space video prediction

We present an example video frame directly generated from the predicted latent $\hat{z}_t$ (without residual compensation) in Figure 1 top right and Figure 5 second left. The result shows that our approach produces effective next-frame prediction to achieve high compression rates. Figure 6 right shows the quality of video entirely generated from the predicted latent $\hat{z}_t$ in terms of PSNR on UVG [35] dataset. For this experiment, the prediction-only mode produces the video sequences from predicted next-frame latents $\hat{z}_t$ without residual compensation whereas the input to the predictor is the reconstructed latents $\tilde{z}_{<t}$. The result in Figure 6 right shows that our approach provides high quality next-frame prediction while the average prediction quality saturates as the bit rate increases. Note that the bit rate for the prediction-only mode is overstated because it is mostly dominated by residuals which are unused in the prediction-only mode. The observation implies that while the reconstructed latents, $\tilde{z}_t$, can learn the spatial domain correlation within a frame very well, there remains non-negligible inter-frame temporal prediction errors caused by complex motion dynamics. In the video compression task, we alleviate this issue and achieve significant quality improvements by saving and transmitting key-frames and residuals. For the majority of *normal* video frame sequences that have strong temporal correlations, the proposed prediction method provides reliable and accurate prediction for compression. For occasional abnormal sequences, we exploit the limitation of the prediction network for the task of abnormal event detection (Sec. 4.5).

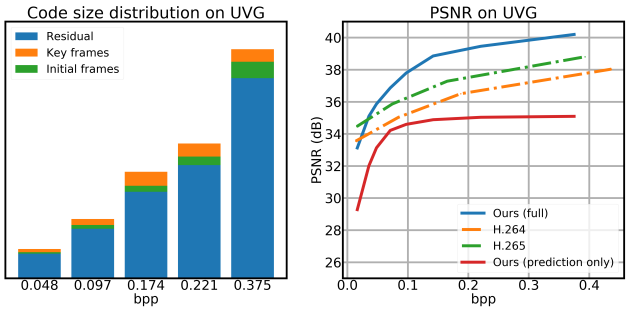

Figure 6. Code length distribution for different compression ratios (left) and evaluation of next-frame prediction performance (right). The prediction-only method generates the video sequences from predicted next-frame latents $\hat{z}_t$ without residual compensation whereas the input to the predictor is the reconstructed latents $\tilde{z}_t$. Note that bpp values for the prediction-only curve do not reflect the actual code size (which should be very close to 0 as residuals are discarded). They are just proportional to the latent space dimension we adopted for the experiment.

4.4. Compressed data size distribution

As described in Section 3.1, we first transmit $k = 6$ *initial frames* without prediction in their latent representation z_{opt}, and they are fed to the prediction ConvLSTM network to initialize the latent sequence. During the regular codec operation after initialization, we keep monitoring the quality of reconstructed frames $\tilde{x}_t$ compared with the raw frame by evaluating the MSE distortion $d(x_t, \tilde{x}_t) = \|x_t - \tilde{x}_t\|^2$. When the distortion exceeds a predefined threshold, we declare that frame a *key-frame* and transmit the BPG-coded frame directly to the receiver (without using the latent domain residual). To further inspect the implication of this proposed approach, we analyze the distribution of code length for different bit-rates in Figure 6 left. The result shows that residual latent vectors dominate the overall compressed data as expected. The proportion of key-frames is dependent on the video content (abrupt scene changes incur more key-frames) but it is less significant especially for very low or high compression rates. We set a lower distortion threshold for the higher image quality target. Thus it is likely to encounter more intra-coded key-frames for higher quality video compression. On the other hand, allowing more bits per pixel reduces the distortion, thereby decreases the num-

ber of bits for key-frame transmission. Because of these two counteracting effects, the bit allocation for the key-frames shown in Figure 6 left shows a non-monotonic pattern with relatively fewer bits at two extremes (lowest and highest rates), while more bits are allocated to key-frames when the compression is at a medium level.

Our evaluation on UVG dataset confirms that the key-frames occupy on average only 8.73% of the total code length. The bit streams for prediction residual encoding account for 84.8% of the total compressed bit sequence on average, while the rest 6.48% contributes to the initial frames. According to our inspection on the composition of the compressed signal, the residual r_t is 8× more sparse (fewer non-zero values) than the target latent representation $z_{opt,t}$ on average. This validates that generating accurate prediction is a key enabler for our codec.

4.5. Application: scene anomaly detection

We extend our study and experiments to anomaly detection on surveillance video clips, which mostly contain homogeneous contents with relatively small changes between scenes. With the assumption that our prediction model reliably predicts the next frame in normal scenes, an anomalous event is detected when the difference between the predicted latent vector and the target z_{opt} of the frame is substantial. We quantify the error of prediction caused by the abnormal event by computing the Euclidean distance between the predicted and target latent vector.

Following the regularity score proposed in [13], $e(t)$ is the L_2 distance between the prediction and target latent at frame index t. The score $S(t)$ reflects the normality of the frame within the sequence of time duration T.

$$e(t) = \|z_{opt} - \tilde{z}_t\|_2$$
$$S(t) = 1 - \frac{e(t) - \min_\tau(e(\tau))}{\max_\tau(e(\tau))}, \quad \tau = t-T, t-T+1, \cdots, t \tag{12}$$

The regularity score (eq.12) indicates that relatively small prediction error produce a score close to 1, while it will drop significantly when large prediction error is encountered because of an abnormal (i.e., unseen during the prediction training) event in the scene. Figure 7 visualizes the change of the regularity score in a video clip.

We benchmark the anomaly detection performance on several widely used datasets with distinctive features. UCSD [32] Ped1 contains 40 abnormal events in 70 video clips, and UCSD Ped2 has 12 abnormal events in 28 videos. The Subway entrance / exit dataset [2] has 96 / 43 minutes video with 66 / 19 abnormal events. Avenue dataset [29] comprises overall 47 abnormal events. Our predictor network is solely trained for the video compression task, and it has not been retrained for the anomaly detection. We evaluate our approach on the test sequences of those datasets in terms of area under the Receiver Operation Characteristic (ROC) curve (AUC), which cumulatively reflects the ROC metric. Generally, higher AUC indicates better performance.

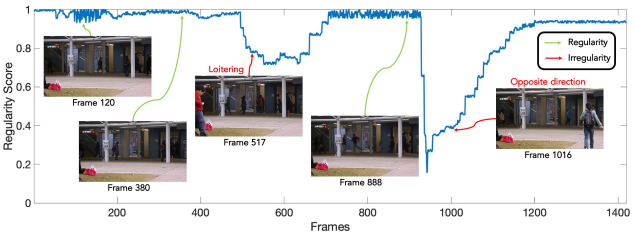

Figure 7. Regularity score (blue curve in the figure) along the frame sequence. Normal scenes typically score > 0.9 whereas abnormal events cause steep score degradation.

Table 1. Anomaly detection performance evaluated by area under ROC curve (AUC %).

Methods	UCSD Ped 1	UCSD Ped 2	Subway Entrance	Subway Exit	CUHK Avenue
Wang *et al.* [47]	72.7	87.5	81.6	84.9	–
Hasan *et al.* [20]	81.0	90.0	**94.3**	80.7	70.2
Chong *et al.* [13]	89.9	87.4	84.7	94.0	80.3
Liu *et al.* [27]	83.1	**95.4**	–	–	84.9
Gong *et al.* [18]	–	94.1	–	–	83.3
Ours	**90.9**	93.6	88.2	**94.5**	**85.4**

In our scheme, normal scenes with learned patterns mostly do not trigger false alarms as they create small fluctuations in the regularity score $S(t)$. However, when an abnormal event happens, the regularity score significantly drops with a high probability. This abnormal event detection is just a byproduct of our compression algorithm. Nonetheless, our event detection performance shown in Table 1 exhibits comparable/superior accuracy compared with others specifically designed for the task.

5. Conclusion

We propose a GAN based framework that accomplishes video prediction and compression. Our method simultaneously learns a transform of the original video into a lower-dimensional latent representation as well as a temporally-conditioned probabilistic model. The performance evaluations show that our work achieves superior/competitive result compared to other (learning-based) codecs for a wide range of rate-distortion trade-off. Our performance gain majorly attributes to the approach that reduces both the spatial and temporal redundancy by combining image compression and video prediction in latent space. We also demonstrate an application using the video prediction score to detect an abnormal event, showing competitive accuracy compared to algorithms specifically optimized for that task.

References

[1] *Video Trace Library*. http://trace.eas.asu.edu/index.html.

[2] A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz. Robust real-time unusual event detection using multiple fixed-location monitors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(3):555–560, 2008.

[3] Eirikur Agustsson, David Minnen, Nick Johnston, Johannes Balle, Sung Jin Hwang, and George Toderici. Scale-space flow for end-to-end optimized video compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[4] Eirikur Agustsson, Michael Tschannen, Fabian Mentzer, Radu Timofte, and Luc Van Gool. Generative adversarial networks for extreme learned image compression. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 221–231, 2019.

[5] Mohammad Babaeizadeh, Chelsea Finn, Dumitru Erhan, Roy H. Campbell, and Sergey Levine. Stochastic variational video prediction. In *International Conference on Learning Representations*, 2018.

[6] Mohammad Haris Baig, Vladlen Koltun, and Lorenzo Torresani. Learning to inpaint for image compression. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 1246–1255. Curran Associates, Inc., 2017.

[7] Johannes Ballé, Valero Laparra, and Eero P. Simoncelli. End-to-end optimized image compression. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

[8] Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston. Variational image compression with a scale hyperprior. In *International Conference on Learning Representations*, 2018.

[9] Fabrice Bellard. *BPG Image Fromat*. https://bellard.org/bpg/.

[10] João Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. *CoRR*, abs/1705.07750, 2017.

[11] T. Chen, H. Liu, Q. Shen, T. Yue, X. Cao, and Z. Ma. Deepcoder: A deep neural network based video compression. In *2017 IEEE Visual Communications and Image Processing (VCIP)*, pages 1–4, 2017.

[12] Zhengxue Cheng, Heming Sun, Masaru Takeuchi, and Jiro Katto. Learning image and video compression through spatial-temporal energy compaction. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[13] Yong Shean Chong and Yong Haur Tay. Abnormal event detection in videos using spatiotemporal autoencoder. *CoRR*, abs/1701.01546, 2017.

[14] Emily Denton and Rob Fergus. Stochastic video generation with a learned prior. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1174–1183, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

[15] Abdelaziz Djelouah, Joaquim Campos, Simone Schaub-Meyer, and Christopher Schroers. Neural inter-frame compression for video coding. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[16] Chelsea Finn, Ian Goodfellow, and Sergey Levine. Unsupervised learning for physical interaction through video prediction. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 64–72. Curran Associates, Inc., 2016.

[17] Adam Golinski, Reza Pourreza, Yang Yang, Guillaume Sautiere, and Taco S. Cohen. Feedback recurrent autoencoder for video compression. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, November 2020.

[18] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[19] Amirhossein Habibian, Ties van Rozendaal, Jakub M. Tomczak, and Taco S. Cohen. Video compression with rate-distortion autoencoders. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[20] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K. Roy-Chowdhury, and Larry S. Davis. Learning temporal regularity in video sequences. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[21] Zhihao Hu, Zhenghao Chen, Dong Xu, Guo Lu, Wanli Ouyang, and Shuhang Gu. Improving deep video compression by resolution-adaptive flow coding. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 193–209, Cham, 2020. Springer International Publishing.

[22] Radford M. Neal Ian H. Witten and John G. Cleary. Arithmatic coding for data compression. *Communications of the ACM*, 30:520–540, 1987.

[23] Nick Johnston, Damien Vincent, David Minnen, Michele Covell, Saurabh Singh, Troy Chinen, Sung Jin Hwang, Joel Shor, and George Toderici. Improved lossy image compression with priming and spatially adaptive bit rates for recurrent networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[24] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1558–1566, New York, New York, USA, 20–22 Jun 2016. PMLR.

[25] Alex X. Lee, Richard Zhang, Frederik Ebert, Pieter Abbeel, Chelsea Finn, and Sergey Levine. Stochastic adversarial video prediction. *CoRR*, abs/1804.01523, 2018.

[26] B. Liu, A. Cao, and H. Kim. Unified signal compression using generative adversarial networks. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3177–3181, 2020.

[27] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. Future frame prediction for anomaly detection - A new baseline. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 6536–6545. IEEE Computer Society, 2018.

[28] Salvator Lombardo, Jun Han, Christopher Schroers, and Stephan Mandt. Deep generative video compression. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 9287–9298. Curran Associates, Inc., 2019.

[29] C. Lu, J. Shi, and J. Jia. Abnormal event detection at 150 fps in matlab. In *2013 IEEE International Conference on Computer Vision*, pages 2720–2727, 2013.

[30] 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 Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part II*, volume 12347 of *Lecture Notes in Computer Science*, pages 456–472. Springer, 2020.

[31] Guo Lu, Wanli Ouyang, Dong Xu, Xiaoyun Zhang, Chunlei Cai, and Zhiyong Gao. Dvc: An end-to-end deep video compression framework. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[32] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos. Anomaly detection in crowded scenes. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1975–1981, 2010.

[33] Michaël Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. In Yoshua Bengio and Yann LeCun, editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.

[34] Fabian Mentzer, Eirikur Agustsson, Michael Tschannen, Radu Timofte, and Luc Van Gool. Conditional probability models for deep image compression. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[35] A. Mercat, M. Viitanen, and J. Vanne. Uvg dataset: 50/120fps 4k sequences for video codec analysis and development. *ACM Multimedia System Conference*, 2020.

[36] David Minnen, Johannes Ballé, and George D Toderici. Joint autoregressive and hierarchical priors for learned image compression. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 10771–10780. Curran Associates, Inc., 2018.

[37] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *CoRR*, abs/1411.1784, 2014.

[38] Debargha Mukherjee, Jim Bankoski, Adrian Grange, Jingning Han, John Koleszar, Paul Wilkins, Yaowu Xu, and Ronald S Bultje. The latest open-source video codec vp9 - an overview and preliminary results. 2013.

[39] Hyunjong Park, Jongyoun Noh, and Bumsub Ham. Learning memory-guided normality for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[40] Ao Ren, Tianyun Zhang, Shaokai Ye, Jiayu Li, Wenyao Xu, Xuehai Qian, Xue Lin, and Yanzhi Wang. ADMM-NN: an algorithm-hardware co-design framework of dnns using alternating direction method of multipliers. *CoRR*, abs/1812.11677, 2018.

[41] Oren Rippel, Sanjay Nair, Carissa Lew, Steve Branson, Alexander G. Anderson, and Lubomir Bourdev. Learned video compression. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[42] S. Santurkar, D. Budden, and N. Shavit. Generative compression. In *2018 Picture Coding Symposium (PCS)*, pages 258–262, 2018.

[43] 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.

[44] Lucas Theis, Wenzhe Shi, Andrew Cunningham, and Ferenc Huszár. Lossy image compression with compressive autoencoders. In *International Conference on Learning Representations*, 2017.

[45] George Toderici, Sean M. O'Malley, Sung Jin Hwang, Damien Vincent, David Minnen, Shumeet Baluja, Michele Covell, and Rahul Sukthankar. Variable rate image compression with recurrent neural networks. In *International Conference on Learning Representations*, 2016.

[46] George Toderici, Damien Vincent, Nick Johnston, Sung Jin Hwang, David Minnen, Joel Shor, and Michele Covell. Full resolution image compression with recurrent neural networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[47] T. Wang and H. Snoussi. Histograms of optical flow orientation for visual abnormal events detection. In *2012 IEEE Ninth International Conference on Advanced Video and Signal-Based Surveillance*, pages 13–18, 2012.

[48] Yilin Wang, Sasi Inguva, and Balu Adsumilli. Youtube UGC dataset for video compression research. *CoRR*, abs/1904.06457, 2019.

[49] 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.

[50] Chao-Yuan Wu, Nayan Singhal, and Philipp Krähenbühl. Video compression through image interpolation. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VIII*, volume 11212 of *Lecture Notes in Computer Science*, pages 425–440. Springer, 2018.

[51] Ren Yang, Fabian Mentzer, Luc Van Gool, and Radu Timofte. Learning for video compression with hierarchical quality and recurrent enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.