IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 9 April 2024, accepted 3 May 2024, date of publication 9 May 2024, date of current version 11 June 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3399174

RESEARCH ARTICLE

A Semantic Communication and VVC Based Hybrid Video Coding System

PRABHATH SAMARATHUNGA[1], (Graduate Student Member, IEEE), YASITH GANEARACHCHI[1], (Graduate Student Member, IEEE), THANUJ FERNANDO[1], (Student Member, IEEE), ADHURAN JAYASINGAM[2], (Member, IEEE), INDIKA ALAHAPPERUMA[1], (Graduate Student Member, IEEE), AND ANIL FERNANDO[1], (Senior Member, IEEE)

[1]Department of Computer and Information Sciences, University of Strathclyde, G1 1XQ Glasgow, U.K.
[2]Wireless Media and Networking Research Group, Faculty of Engineering, Computing and the Environment, Kingston University, London, KT1 2EE Kingston upon Thames, U.K.

Corresponding author: Anil Fernando (anil.fernando@strath.ac.uk)

This work was supported in part by the Global Research Scholarships of the University of Strathclyde, Glasgow.

ABSTRACT Requirements of next-generation video applications are becoming a challenge for conventional video coding systems, although they have evolved over decades to accommodate the most demanding of current video applications. Semantic communications, built on the concept of transmitting just the semantics of a message and allowing the receiver to reconstruct the message based on a shared context, is a non-conventional approach being considered to overcome these challenges and improve performance of video coding systems. In this paper, a first such semantic communication-based video coding system in hybrid mode is proposed, which uses an autoencoder-based semantic encoder for inter coding, augmented by the intra coding capabilities of Versatile Video Coding (VVC) to encode key frames that form the context for the semantic communication and the residuals for improving the fidelity of the output frames. For a range of videos with differing levels of complexity, the proposed system consistently outperforms High Efficiency Video Coding (HEVC) and Advanced Video Coding (AVC) in terms of rate distortion metrics quantified by Bjontegaard Delta Rates. It also outperforms Versatile Video Coding with videos with low or high complexity, but slightly falls behind with videos with medium complexity, which can be improved by addressing the open research areas that stem from this work. The proposed system demonstrates the potential of semantic communication based video coding systems to consistently outperform state-of-the-art conventional video coding systems over a wide range video applications.

INDEX TERMS Autoencoders, deep neural networks, semantic communication, video coding, video communication.

I. INTRODUCTION

New media formats such as ultra high definition (UHD) television, high dynamic range (HDR), extended reality (XR), omnidirectional video, and interactive storytelling are now widely supported by consumer devices. This has led to a rapid growth in the volume of video content distributed over the Internet and other access media, accounting for more than 60% of the current mobile network traffic. This share is expected to increase to 80% by 2028 [1], making it increasingly challenging to ensure a consistent end-to-end quality of experience (QoE) of video during its production, storage, and distribution, especially due to video characteristics, such as bit depth, color rendition, frame rate, and resolution, which also increase on top of the volume.

Raw video footage is large in size and would not be very practical to use without video compression (video coding) to make it small enough to store and transmit within the resource limitations of storage and transmission media. Video coding standards define encoded bitstream and decoder

The associate editor coordinating the review of this manuscript and approving it for publication was Muhammad Sharif.

specifications to enable widespread use of compressed video with a wide range of device and software platforms.

The video coding landscape has been shaped by a series of standards developed by the International Telecommunications Union Telecommunication Standardization Sector (ITU-T) and the Moving Picture Expert Group (MPEG) of the International Standards Organization (ISO). The most popular video coding standards from these organizations are H.264 Advanced Video Coding (AVC)/MPEG-4 part-10 standardized in 2003, and H.265 High Efficiency Video Coding (HEVC)/MPEG-H part-2 standardized in 2013. H.266 Versatile Video Coding (VVC)/MPEG-I part-3 standardized in 2020 is the latest standard in this family and represents the state-of-the-art in video coding. However, there are other popular video coding formats that include the Audio Video Coding Standard (AVS), video codec 1 (VC-1), Windows Media Video 9 (WMV-9), Theora, VP8, VP9, Daala and AOMedia Video 1 (AV-1) [2].

These mainstream video coding standards are based on statistical signal processing (SSP) techniques to exploit statistical redundancies and quantization to exploit perceptual redundancies. However, with increasing bit depth, frame rate, and resolution of the video, the complexity and computational workload of the coding algorithms inevitably increase in part due to the linearity of the transformations used to compress the data, resulting in an increase in the time taken to encode and decode the video, as demonstrated by [3] and [4]. Therefore, while still being well capable of providing sufficient QoE for current video applications such as UHD and HDR content, even state-of-the-art video coding standards will face significant challenges in delivering the next generation of video applications where resolutions beyond 16K (15,360×8,640 resolution) and frame rates beyond 300 frames per second are being considered. In response, the use of deep neural networks (DNNs) that allow the use of nonlinear transformations to achieve compression in video has been explored as an alternative with some positive results [5]. Moreover, there are inherent challenges associated with them, such as transmission of the DNN from the transmitter to the receiver and training of the DNN, which have hindered their development towards becoming mainstream coding standards.

Inspiration for a novel approach to video communication that can circumvent the channel capacity challenge faced by conventional video coding systems can be found in one of the first concepts of information theory. Shannon and Weaver [6] suggest that the desired effect of communication can be achieved by successful extraction and transmission of just the *semantics* of a message, which can be text, images or videos. When the context of semantic extraction is shared between the transmitter and the receiver, the semantic can even be used to reconstruct the original in the receiver. This is analogous to a human using written word to describe an event, and a reader who is sufficiently fluent in the language used for writing being able to visualize the event based on the description, where the written text constitutes the semantics of the event, and the vocabulary and grammar of language constitute the shared context.

While the concept of semantic communication has been around for decades, a technique to efficiently reduce a message to its semantic form was not feasible due to the inaccessibility of the appropriate techniques and computational power required for implementation. However, recent advances in artificial intelligence (AI) and machine learning (ML) make it possible to implement semantic extraction and semantic-based reconstruction using DNN, such as autoencoders, which has generated renewed interest in semantic communications [7].

A semantic communication system (Fig. 1) consists of a semantic encoder and decoder that convert the message to semantics at the transmitter and the conversion of semantics back to the message at the receiver, both using the same context (or common knowledge) to perform encoding and decoding. In implementations, how the context is interpreted by the encoder may not always be identical to how it is done by the decoder, adding semantic noise to the end-to-end communication process on top of the omnipresent electronic noise. Although this may not affect the effectiveness of the communication, it does reduce the fidelity of the output when attempting to reconstruct the message based on semantics. This is a challenge which semantic communication systems, especially when used for media transmission, must address.

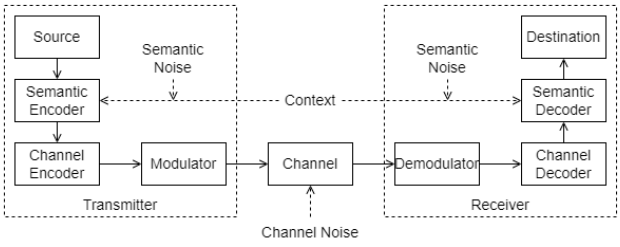


FIGURE 1. Typical semantic communication system.

In response, we propose a hybrid video coding system that combines semantic communications and an SSP based video coding system (VVC) to achieve rate distortion performances on par with those of state-of-the-art conventional video coding systems. An autoencoder (AE) [8] is chosen to perform semantic encoding and decoding, and is trained for each group of pictures (GOP) to minimize the mean square error (MSE) between the input frames and the output frames, effectively making the latent vector in the bottleneck layer a semantic representation of each frame. A frame decoder at the receiver is then trained using the AE decoder layers to minimize MSE between the key frames (first frame of each GOP) shared by the transmitter and the frame reconstructed using the latent vector received from the transmitter. The AE decoder layers on the transmitter (semantic encoder) and the AE decoder layers on the receiver (semantic decoder) have the same structure and are initialized with the same seeds to further optimize the performance, while it becomes no longer necessary to use the same trained decoder layers from the

transmitter at the receiver. Therefore, the structure of the AE layers, the seed used for initialization, and the key frames represent the *context* of the semantic video compression system, while the latent vector represents its *semantics*.

The key frames are encoded using the state-of-the-art intra coding capabilities of VVC. In addition, to overcome the challenge of reconstructing high-fidelity frames, a residual is extracted by comparing the original frame with the reconstructed frame (using a semantic decoder similar to that used at the receiver at the transmitter). This residual is encoded using VVC and sent to the receiver, which is added to the reconstructed frames at the receiver to produce a high-fidelity output image. The use of a SSP based codec (VVC) to perform context encoding and residual encoding along with the use of semantic communication concepts to extract *semantics* of each frame makes the proposed system a hybrid of semantic communications and SSP.

Performance of the proposed system is benchmarked against state-of-the-art video codecs based on VVC, HEVC, and AVC by comparing the rate distortion performances based on the peak signal-to-noise ratio (PSNR), structural similarity index (SSIM) and video multi-method assessment fusion (VMAF) metrics. The comparative performance of the proposed system is further quantified using the Bjontegaard Delta (BD) rates calculated for each metric.

The key novel contributions that we make to the field of video coding through this paper are:

- pioneering in an end-to-end hybrid video compression system based on the concept of semantic communications for inter coding and VVC for intra coding and residual compression.
- demonstrating a novel mechanism for remote training of a semantic communication based frame decoder, allowing the proposed framework to function without the necessity of transporting the decoder parameters from the transmitter to the receiver, which is the first such attempt to use a remotely trained DNN based system for video encoding for transmission.
- demonstrating the performance of the proposed system with a range of videos with differing structural information (SI) and temporal information (TI) content and benchmarking it against state-of-the-art codecs (VVC, HEVC, and AVC) in terms of rate distortion performance quantified using BD rates.

The remainder of this paper is organized into four sections. Section II provides a brief overview of video coding using conventional video coding and DNN-based video coding, as well as the current state of research on semantic communications. Section III describes the proposed semantic and VVC based hybrid video codec and the experiments carried out to evaluate its rate distortion performance against VVC, HEVC, and AVC. Section IV presents the results of the experiments and a brief discussion on their impact on the future of video coding, leading to Section V which summarizes the work and a brief note on how to further develop the proposed system to meet the emerging requirements of the media, communications, and consumer electronics industries.

II. RELATED WORK

Conventional video coding has evolved over four decades from the earliest digital video compression standard ratified in 1984, H.120, to the state-of-the-art standard ratified in 2020, VVC. However, the efficiency of conventional video coding techniques are fast reaching their limits, and alternative approaches for video coding such as DNN based video coding have been explored in parallel as potential candidates for the next generation of video coding. Semantic communication, which has started to gain interest in recent years, offers a novel paradigm which could be used for video coding and transmission in multiple applications where conventional video coding systems cannot achieve any further significant coding gain.

A. CONVENTIONAL VIDEO CODING

Early video formats used in transmission and storage systems were analog and aimed primarily at television broadcast. Videoconferencing over public telephone networks was the first application of digital video that required the creation of standardized digital video formats [9]. The first digital video standards ratified by ITU-T included H.120 (version 1) [10], and H.261 which was the first widely adapted digital video coding standard [11]. ISO's MPEG introduced MPEG-1 video in 1993, following which it joined with the Video Coding Experts Group (VCEG) of ITU-T to release subsequent standards H.262 and H.263.

VCEG and MPEG established the Joint Video Team (JVT) in 2001, which resulted in the publication of the AVC or H.264 standard in 2004 (technically identical to MPEG-4 part 10), which eventually became the most widely adapted video coding standard. This was followed by HEVC or H.265 (also ratified as MPEG-H part 2), in 2013 [12] providing a coding gain of 50% over AVC [13], [14]. The state-of-the-art video compression standard, VVC or H.266 (also ratified as MPEG-I part 3), followed in 2020 and the development of VVC codecs is currently gaining momentum with an expected coding gain of 50% over HEVC along with support for a much wider range of industry applications [12]. However, starting from AVC, the H.26X family of video coding standards is distributed under a royalty-based licensing scheme, which limits its appeal to large-scale video distribution and streaming services.

Therefore, there are a multitude of alternative video coding standards and formats published by entities ranging from major industry players to industry consortiums that are proprietary for specific applications or are royalty free for industrial use. AVS [15], WMV [16], Theora [17], VP8 [18], VP9 [19], and AV1 [20] are all such video compression formats that have a significant market presence.

All these standards are based on SSP and therefore make use of several key techniques that include block partitioning,

motion compensation or interpicture prediction, intrapicture prediction (intra coding), transformation (typically in the frequency domain and commonly using the discrete cosine transform), quantization, entropy coding, and in-loop filtering [21]. HEVC added more flexibility and effectiveness to these concepts introduced in AVC, by adding features such as range extensions (RExt) to allow modification of the video format, improvements in coding efficiency, precision, and throughput optimizations, scalable HEVC extensions (SHVC) to extend temporal scalability to spatial, quality, bit depth, and color range scalability, multiview (MV-HEVC) and 3-D extensions (3)-D-HEVC) combining multilayer design with scalability features and screen content coding (SCC) extensions [13]. VVC, again based on the coding tools used in HEVC, includes a further improved and refined set of techniques that make it even more flexible and adaptable for future video applications, albeit at the cost of a general increase in complexity [11].

However, considering the computational complexity and optimizations that VVC entails, its successor is expected to face significant challenges in realizing better rate distortion performance using conventional techniques alone [12]. This is leading to greater interest in alternative video compression methods, such as AI-based systems, improved perceptual quality measurement systems to optimize quality-based compression, and video coding for machines, which does not require full reconstruction of the videos to be effective in communication.

B. DEEP NEURAL NETWORK BASED VIDEO CODING

DNNs work in a similar way to human perception, using multiple layers of nodes that are interconnected with different weights and activated based on a range of input-output functions set on different biases. The strength of DNNs lies in its ability to identify complex patterns in the input data which are impossible to do using fixed rules, as done in conventional signal processing algorithms.

Since late 1980s a range of DNNs has been explored for image compression, which has formed the basis for using them for video compression. The DNNs used for image coding have evolved from multilayer perceptron (MLP) based image coding, random neural network based image coding, convolutional neural network (CNN) based image coding, recurrent neural network (RNN) based image coding, generative adversarial network (GAN) based image coding, and more recently, vision transformer based image coding [22]. Extending these developments to video coding has been an active area of research over the past decade, especially since the introduction of HEVC, and DNN-based improvements to modules within HEVC and other mainstream video coding standards were one of the first adoptions of AI/ML for video coding. DNNs were used to enhance and enhance the tools used for intra prediction, inter prediction, quantization, entropy coding, and loop filtering. New DNN-based video coding frameworks have also been introduced, using CNNs, Voxel CNNs, and long short-term memory (LSTM) encoder-decoder frameworks. Key challenges in using DNN-based video coding include semantic fidelity oriented video compression, rate distortion optimization guided training and adaptive switching, and efficient memory and computation design [22].

One of the first end-to-end DNN-based video compression models was deep video compression (DVC), which jointly optimized the motion and residual information of a video by implementing two AE style DNNs jointly trained on a single loss function to carry out motion and residual compression. DVC was demonstrated to have better PSNR performance compared to AVC and reached HEVC performance in terms of the multiscale structural similarity index method (MS-SSIM) [23]. Further improvements on DVC using a DNN based motion prediction and refinement network that uses intra-frames (I-frames) and predicted frames (P-frames) built on a motion predictor-net for tracking motion using differential motion vectors and a refinenet [24] demonstrated competitive rate distortion performance over HEVC class B, class D [13] and Ultra Video Group (UVG) [25] datasets in terms of MSE and MS-SSIM. DVC adds additional complexity to the framework due to the attempt to use DNNs for residual coding and motion estimation, and the performance was only benchmarked for AVC and HEVC. It has not been designed targeting video transmission.

A new video compression architecture using feedback recurrent AE for real-time video compression adopting the approach taken by conventional video codecs such as AVC and HEVC [26] demonstrates comparable SSIM and compression rate performance with AVC and HEVC for those commonly used in streaming applications. But the system faces challenges in temporal consistency and color shift, especially when the size of the GOP is increased.

A three-dimensional AE featuring a discrete latent space and an auto-regressive prior was evaluated against AVC and HEVC, exhibiting some advantages in terms of adaptive compression compared to conventional codecs [27]. However, it is not benchmarked against VVC and the applications of the proposed method are for domain-specific videos, such as specific industrial applications, and it depends on the trained decoder being available at the receiver, which requires the use of a neural network compression and transmission method essential for any practical implementations.

Although the performance of the PSNR was not better than the default HEVC settings [28] proposes a spatial-temporal video compression network (STVC) using spatial-temporal priors with an attention (STPA) module that is based on AE architectures, which demonstrates better rate distortion performance than AVC and DVC. However, STVC also suffers from the added complexity and latency of using DNN techniques to achieve residual coding.

A Video compression system that used a frame AE, flow AE, and motion extension network (Flow-MotionNet) [29] delivered improved performance compared to AVC and HEVC in terms of PSNR and SSIM, but with a higher

processing time per frame (TPF). An optical flow residual coding method [30] specifically for videos that have a strong interframe correlation, such as surveillance video or teleconferencing video, demonstrated quality improvements of 1.2 dB compared to DVC. However, the system is not benchmarked against VVC, and furthermore it is not investigated with general video applications, limiting its effectiveness to a specific domain.

In general, the utilization of AEs for video compression has shown encouraging rate distortion results compared to conventional video codecs such as AVC and HEVC, but with a trade-off of increased computational complexity, as evidenced by longer processing times. However, their performance still falls short of achieving the rate distortion performance of HEVC codecs over a range of videos, and there has not been any AE-based framework put forth that can match the rate distortion performance of latest VVC codecs, and as such no efforts have been made yet.

C. SEMANTIC COMMUNICATIONS FOR VIDEO TRANSMISSION

Semantic communications was first discussed in the introduction of the mathematical theory of communication, where communication is defined as a problem at three levels: technical level, semantic level, and effectiveness (or influence) level [6]. The accuracy of the transfer of information (in the form of a bit stream) from the sender to the receiver is the primary concern of the technical problem and is inherent in all forms of communication. Going beyond the accuracy of the communicated bit stream, the semantic problem concerns itself with the meaning of the information interpreted by the receiver in comparison to the meaning of the information intended by the sender. The third problem of effectiveness is itself concerned with the success of which the meaning understood by the receiver leads to the action that the sender expected [31].

The development of communications focused mainly on solving the technical problem of communications and exploring the limits of information capacity that can be achieved with the two constraints of the signal-to-noise ratio (SNR) and the channel bandwidth. However, with the advent of DNNs, it has become possible to develop systems where senders can extract the meaning of a message and transmit it for a receiver to infer the meaning with high accuracy. Since the information content of the meaning, or *semantic*, of a message is significantly less than the information content of a message encoded as a bit stream, this creates the possibility of transmitting a larger amount of information within a restricted bandwidth. Furthermore, with the increased prevalence of Machine-to-Machine (M2M) communication and the proliferation of the Internet of Things (IoT), accurate communication of the meaning of a message even when the receiver does not have a fully accurate bit stream has become more relevant. If the sender and receiver share prior knowledge or context of the communication, the resulting semantic communication system is capable of significantly improving the throughput of a channel [7]. Building on these concepts, contemporary research shows that semantic communications are a viable means to preserve semantic information while compressing information to considerably small sizes, even in the presence of intentional data loss at the bit level compared to conventional encoding methods [32], [33], [34], [35], [36].

The first discussions on semantic communication by [6] were based on text transmission, which was more recently explored in detail using a range of deep learning techniques [7]. DeepSC is a semantic communication based text transmission system that attempts to maximize capacity and minimize semantic errors by accurately recovering *semantic* instead of symbol level or *technical* problem, and demonstrates high fidelity of the semantic communication system in low signal-to-noise ratio (SNR) channels [37], [38]. Similar observations have also been made on higher compression ratios and better resilience to noise from semantic communication systems in speech transmission [39]. A key challenge identified in these was the introduction of semantic noise to the communication process, which was mainly the result of mismatches in the context referred to by the transmitter and receiver to decode the semantically encoded message.

Extending the concept of semantic communication for image transmission has been attempted using a variety of DNN architectures. An image transmission system with joint source and channel coding implemented using autoencoders was shown to provide a high level of compression while still maintaining semantic fidelity at low SNR compared to conventional channel coding systems [40], and multiple other domain-specific systems based on the same concept used have been studied, such as [41]. GANs have also been used to capitalize on their ability to derive semantic images at the receiving end based on the encoded message, which resulted in highly compressed and noise-resistant image transmission systems [42], [43].

Semantic communication based video communication systems for specific applications, such as video conferencing [44] and video surveillance [45], have also been explored. However, the use of semantic communications for implementing a video coding system that can replace conventional video coding has not yet been attempted, and this work is the first of its kind to explore how state-of-the-art capabilities from semantic communications can be combined with conventional video coding techniques to provide an effective and robust hybrid video coding system.

D. CHALLENGES AND OPPORTUNITIES

To summarize, a key challenge faced by video coding standards is the imminent explosion of system complexity when dealing with the next generation of video applications. This could make obtaining a significant performance improvement over the current generation of video coding

standards by the next exceptionally challenging by continuing to use SSP alone as the basis for video compression. Semantic communications, driven by advances in DNN, may present a solution to reduce the bit rates required for video communication by making use of video *semantics* and *context* for encoding. But in their current state, they alone cannot produce decoded video that will always be acceptable for human viewing, although they may be sufficient for most machine vision-type applications. However, it should be appreciated that conventional video coding standards have evolved over three decades with an immense research effort to become highly efficient in tracking and predicting frames over the temporal dimensions, which cannot be expected to be matched immediately by an emerging technique.

This opens up areas of research on how best to exploit the temporal correlation between frames in semantic communication based video coding systems to match and improve the capabilities currently available in SSP based video coding systems. Meanwhile, there are opportunities to take advantage of the proven features of both systems by exploring hybrid approaches. Section III describes the first attempt to develop and demonstrate a hybrid video coding system using an AE-based semantic communication system augmented by VVC.

III. PROPOSED SYSTEM

The proposed hybrid video coding system uses semantic communication principles, with *semantic* encoding and decoding performed using the AE encoder and decoder layers, respectively, and *context* encoding and decoding performed using VVC. The challenges discussed in Section II-D related to the fidelity of the output frames are addressed by extracting the residual between the original and semantically predicted frames and encoding and decoding them using VVC intra coding, and adding the residual back to the semantically reconstructed frame to generate the output. The proposed system, shown in Fig. 2, is composed of four main components: a semantic encoder, a semantic decoder (frame predictor), a context encoder/decoder and a residual encoder/decoder. The efficient implementation of the proposed system is enabled by a novel approach to remotely train the receiver-side semantic decoder.

The proposed system is tested with 4:2:0 chroma subsampling (YUV420) and on a GOP basis, with the same configuration used for VVC, HEVC, and AVC used as references to compare its performance. It also uses scene transition detection to trigger new GOPs when there is a scene change in the video, making it usable across multiple video applications without being restricted to a particular type of video, such as surveillance video or conference video.

A. SEMANTIC ENCODER

Semantic encoding is the process of extracting the *semantics* of a frame so that it can be used to reconstruct the frame using a compatible semantic decoder. This is typically achieved using a DNN, and in the proposed system, the encoder layers of an AE, shown in Fig. 3, are used to perform semantic encoding to reduce each frame to a latent vector with dimensions 8×1.

The AE is designed to accept a frame with pixel width W, pixel height H and color depth C (which equals 1.5 due to the frame being converted to YUV420 prior to semantic encoding). The tensor representing each frame is flattened to a vector of dimensions $(W \times H \times 1.5) \times 1$ to form an input layer that is fully connected to a hidden layer with dimensions 1024×1 using a Leaky Rectified Linear Unit activation function (LeakyReLU). The hidden layer is fully connected to a bottleneck layer of dimensions 8×1 again using LeakyReLU, from which the latent vector is extracted. To enable AE training using the input frames themselves, the decoder layers are implemented by fully connecting a hidden layer of dimensions 1024×1 with the bottleneck layer using LeakyReLU, which in turn is fully connected to an output layer of dimensions $((W \times H \times 1.5) \times 1$ using a Sigmoid activation function. The output layer is then reshaped to reconstruct the frame with the original dimensions $W \times H \times 1.5$. A learning rate or nonlinearity factor(α) of 0.01 used for training the model to minimize MSE between the input and output frames.

Since the main objective of the proposed system is to reconstruct the original frame using the latent vector on the semantic decoder, it does not necessarily need to be understood by humans. Therefore, the size of the latent vector is optimized to have the minimum possible size after observing the effects of varying the dimensions of the latent vector between 1×1 ($2^0 \times 1$) and 256×1 ($2^8 \times 1$) with the average PSNR of the output frames compared to the input frames as shown in Fig. 4 using input frames with spatial resolution 320×180. The latent vector size of 8×1 was chosen as there is no significant gain in the average PSNR if it is increased further. The semantic encoder training is carried out on a GOP basis, with all frames of the GOP used as training data. For each GOP, the AE is initialized using a Glorot uniform initializer [46] with a preset seed value, which will ensure that the ground state (initial random weights) of the AE will remain the same for each GOP, which is critical to the success of the remote decoder training required to implement the semantic decoder (Section III-B).

The number of training epochs is set at 250 with the intention of overfitting the model to a given GOP, as the differences between frames are minimal for a given GOP, and any scene changes are identified by scene transition detection (Section III-E) to trigger a new GOP.

The model is optimized by varying the structure and depth of the layers to achieve an optimum configuration of one hidden layer of dimensions 1024×1 each for the encoder layers and the decoder layers of the AE. Once the AE training is completed for each GOP, each frame is again fed to the semantic encoder, and the resulting latent vectors are used as the *semantics* of each frame, which is decoded by the semantic decoder (frame predictor) as described in Section III-B.

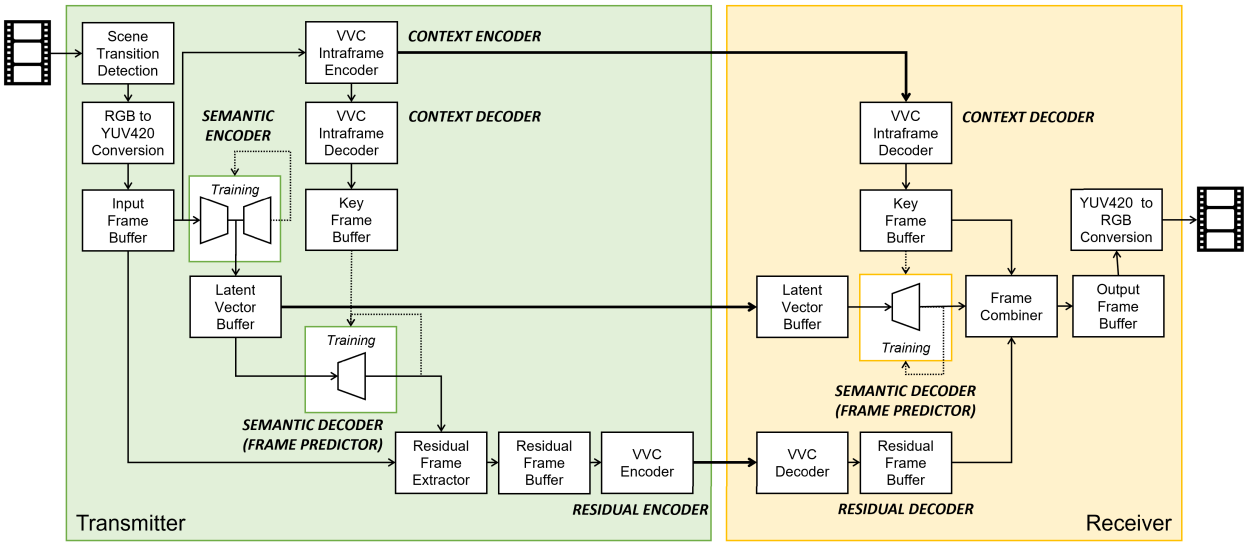


FIGURE 2. Proposed semantic communication and VVC based hybrid video codec.

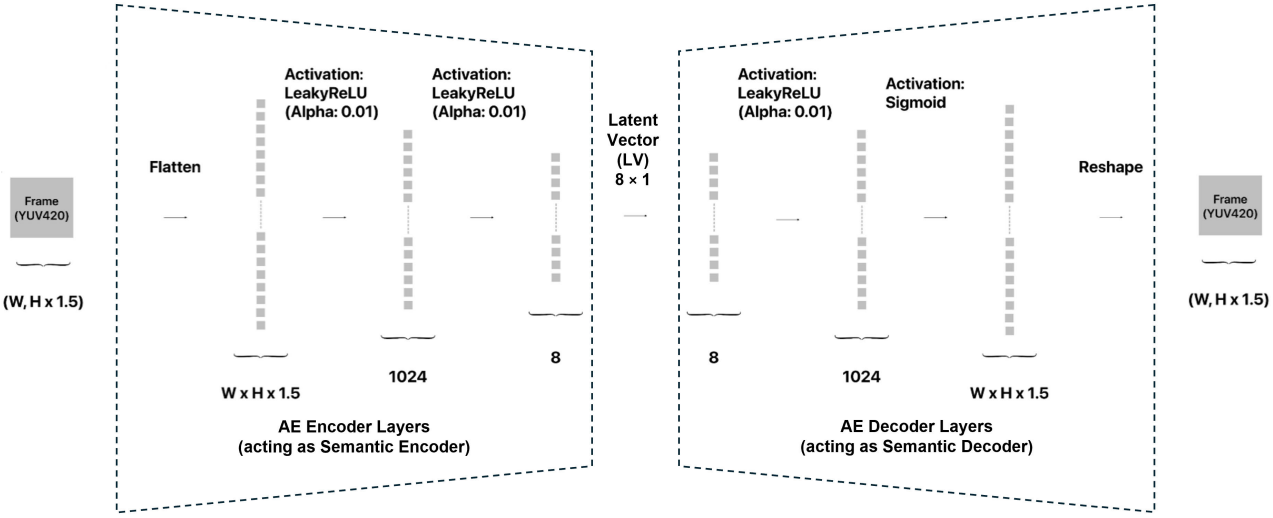


FIGURE 3. AE based architecture used in the semantic encoder and decoder.

B. SEMANTIC DECODER (FRAME PREDICTOR)

The semantic decoder in the proposed system plays a similar role to a frame predictor on a conventional SSP based video coding system. It is implemented using the same structure as the AE decoder layers in the semantic encoder (Section III-A) shown in Fig. 3. However, rather than training the semantic decoder together with the semantic encoder and then moving the AE decoder layers to the receiver, which is the approach taken by studies of other DNN based video transmission systems that have been proposed, as discussed in Sections II-B and II-C respectively. The proposed system uses a novel approach to remotely train the decoder using the *context* provided by the context encoder/ decoder (Section III-C). This enables the decoder at the receiver to be trained for each GOP, as done in the semantic decoder layers at the transmitter, without a significant increase of the computational complexity of the system, while avoiding the requirement of transmitting a large amount of metadata which would be required if the hyperparameters of the AE decoder layers were to be transmitted between the transmitter and receiver for each GOP or scene change.

While the semantic encoder is trained by comparing the input frame and the output frame of the AE, the remote semantic decoder only has access to the key frames decoded through the context decoder and the latent vectors received from the semantic encoder. Therefore, the semantic decoder is trained to optimize the MSE between the key frame and the reconstructed frame from each latent vector. Since overfitting

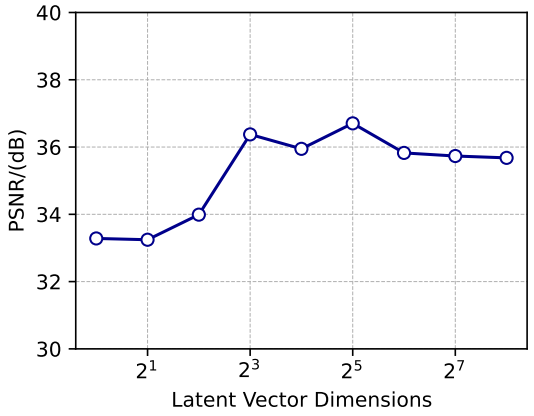


FIGURE 4. Effect of latent vector dimensions on the average PSNR of the reconstructed frame.

the model will lead to higher MSE in this scenario, the training is only carried out for 100 epochs at the semantic decoder, in contrast to the 250 epochs at the semantic encoder.

The effectiveness of remote training of the AE decoder layers can be mathematically verified, considering a video clip with spatial dimensions $W \times H$, where W and H denote the width and height of the frame in pixels, respectively. The number of color channels is denoted by C, which will be equivalent to 3 for the RGB color format, 1.5 for the YUV420 color format, and 1 for the grayscale color format. Therefore, each frame of the video can be represented by a tensor of dimensions $W \times H \times C$, and can be flattened to be represented by a vector $\mathbf{x} \in \mathbb{R}^{WHC \times 1}$.

When the latent vector of the AE is represented by $\mathbf{h} \in \mathbb{R}^{p \times 1}$ where $p \times 1$ is the size of the latent vector and $p \ll WHC$, encoder weight vectors are denoted by $\mathbf{W_{enc}} \in \mathbb{R}^{WHC \times p}$, encoder bias vector is denoted by $\mathbf{b_{enc}} \in \mathbb{R}^{WHC \times 1}$ and an encoder activation function is $f(.)$ the relationship between each frame and its latent vector can be represented by (1).

$$\mathbf{h} = f(\mathbf{W_{enc}}\mathbf{x} + \mathbf{b_{enc}}) \tag{1}$$

The reconstructed output vector from the decoder network, $\hat{\mathbf{x}} \in \mathbb{R}^{WHC \times 1}$, has the same dimensionality as the input ($\mathbf{x}$), and relates to the latent vector ($\mathbf{h}$) through decoder weight vectors denoted by $\mathbf{W_{dec}} \in \mathbb{R}^{WHC \times p}$, an decoder bias vector denoted by $\mathbf{b_{dec}} \in \mathbb{R}^{WHC \times 1}$ and an activation function denoted by $f(.)$ as shown in (2).

$$\hat{\mathbf{x}} = f(\mathbf{W_{dec}}\mathbf{h} + \mathbf{b_{dec}}) \tag{2}$$

The MSE loss function ($L(.)$) between the original input vector and the reconstructed output vector can then be represented by (3) where x_k and $\hat{x}_k$ denote the kth elements of the input and reconstructed output vectors and $N = WHC$.

$$L(\mathbf{x}, \hat{\mathbf{x}}) = \frac{1}{N}\sum_{i=1}^{N}(x_{ki} - \hat{x}_{ki})^2 \tag{3}$$

Conventionally, the encoder and decoder components of the AE are trained simultaneously to minimize the common loss function shown in (3), and would apply when there is no effect of noise on the weights and biases of the decoder component when compared with the encoder.

A situation where a noisy decoder has deviated weights and biases can be represented by (4), where $\mathbf{y} \in R^{WHC \times 1}$ represents the reconstructed output, $\overline{\mathbf{W}}_{\mathbf{dec}} \in R^{WHC \times p}$ is the weight matrix of the decoder at the receiver and, $\overline{\mathbf{b}}_{\mathbf{dec}} \in R^{WHC \times 1}$ is the bias vector of the decoder at the receiver, $\mathbf{h} \in R^{p \times 1}$ is the latent vector transmitted from the transmitter.

$$y = f(\overline{\mathbf{W}}_{\mathbf{dec}}\mathbf{h} + \overline{\mathbf{b}}_{\mathbf{dec}}) \tag{4}$$

Assuming that the noise between the weights and biases of the two sides can be approximated to normally distributed additive white Gaussian noise (AWGN) with zero mean and the variance between $\mathbf{W_{dec}}$ and $\overline{\mathbf{W}}_{\mathbf{dec}}$ is σ_w, and the variance between $\mathbf{b_{dec}}$ and $\overline{\mathbf{b}}_{\mathbf{dec}}$ is σ_b, the relationships between each of these pairs can be represented by (5) and (6).

$$\overline{\mathbf{W}}_{\mathbf{dec}} = \mathbf{W_{dec}} + N(0, \sigma_w^2) \tag{5}$$

$$\overline{\mathbf{b}}_{\mathbf{dec}} = \mathbf{b_{dec}} + N(0, \sigma_b^2) \tag{6}$$

A MSE loss function similar to (3) can now be defined between elements of $\mathbf{x}$ and $\mathbf{y}$ as shown in (7).

$$L(\mathbf{x}, \mathbf{y}) = \frac{1}{N}\sum_{i=1}^{N}(x_{ki} - y_{ki})^2 \tag{7}$$

The accumulated error between $\hat{\mathbf{x}}$ and $\mathbf{y}$, denoted by ϵ, can then be derived as shown in (8) and (9).

$$\epsilon = \mathbf{y} - \hat{\mathbf{x}} \tag{8}$$

$$\epsilon = f(\overline{\mathbf{W}}_{\mathbf{dec}}\mathbf{h} + \overline{\mathbf{b}}_{\mathbf{dec}}) - f(\mathbf{W_{dec}}\mathbf{h} + \mathbf{b_{dec}}) \tag{9}$$

When decoder training is performed to minimize the loss in ϵ, by combining the transmitter side encoder and the receiver side decoder during training, the two activation functions will converge as shown in (10) and (11).

$$\lim_{\epsilon \to 0} f(\overline{\mathbf{W}}_{\mathbf{dec}}\mathbf{h} + \overline{\mathbf{b}}_{\mathbf{dec}}) - f(\mathbf{W_{dec}}\mathbf{h} + \mathbf{b_{dec}}) = 0 \tag{10}$$

$$\therefore f(\overline{\mathbf{W}}_{\mathbf{dec}}\mathbf{h} + \overline{\mathbf{b}}_{\mathbf{dec}}) \approx f(\mathbf{W_{dec}}\mathbf{h} + \mathbf{b_{dec}}) \tag{11}$$

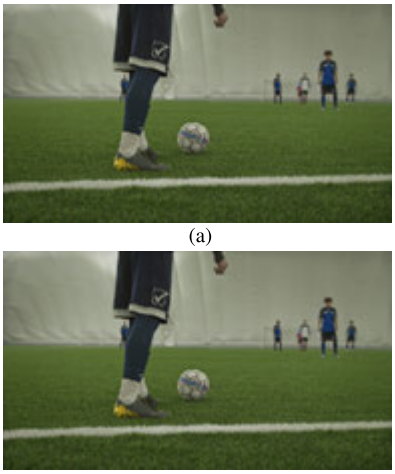
This approximation can be further improved by ensuring that the transmitter AE encoder and decoder layers and the receiver AE decoder layer are initialized using the Glorot uniform initialization (or the Xavier uniform initialization) [46] with the same seed value used by the semantic encoder. This ensures that the ground state of both AE networks is equivalent and will further minimize ϵ derived in (9), leading to a nearly identical reconstruction of the output frame from the transmitter AE decoder layers at output from the receiver AE decoder layers, as shown in Table 1 for a sample of seven different video sequences of varying SI and TI content between medium and high.

This can be further verified by considering the visual reconstructions from an example video with high SI and high

TABLE 1. Comparison of average PSNR of reconstructed videos at transmitter and receiver side semantic decoders.

Video	Average PSNR at Transmitter	Average PSNR at Receiver
Video A	39.688485 dB	39.688492 dB
Video B	39.787627 dB	39.787620 dB
Video C	39.781242 dB	39.781242 dB
Video D	39.758808 dB	39.758808 dB
Video E	39.846144 dB	39.846144 dB
Video F	39.853455 dB	39.853447 dB
Video G	39.773289 dB	39.773296 dB

TI content, as shown in Fig. 5. The frames reconstructed by the transmitter AE decoder layers (Fig. 5a) and by the remotely trained receiver AE decoder layers (Fig. 5b) are not visually discernible from each other. This process is a key innovation of the proposed system and allows it to be adopted in a range of videos without becoming restricted to a specific application such as conferencing or surveillance.



(a) (b)

FIGURE 5. Frames reconstructed at (a) transmitter AE decoding layers and (b) receiver remotely trained AE decoding layers for an example video with high SI and high TI.

The general applicability of remote training of AE decoder layers is further verified when the effect of the GOP size on the average PSNR and the average SSIM of the videos is considered, as shown in Table 2. It demonstrates that remote training of the AE decoder layers at the receiver is consistently able to produce outputs which are equivalent to the AE decoder layers at the transmitter, regardless of the GOP size used. Therefore, the margin of error between the predictions made by the semantic decoder at the receiver using a remotely trained AE is very small and can be neglected. Furthermore, since a GOP based training approach is used, propagation of any impact of remote training is not carried forward beyond a given GOP.

A key factor enabling remote training to be successful is the precise sharing of the *context* between the transmitter and receiver in the form of key frames which is described in Section III-C.

TABLE 2. Comparison of average rate distortion performance of outputs of transmitter side (TX) and receiver side (RX) AE for different GOP sizes.

GOP Size	PSNR(TX)	SSIM(TX)	PSNR(RX)	SSIM(RX)
8	20.27 dB	0.759	20.27 dB	0.759
16	26.64 dB	0.844	26.64 dB	0.844
32	35.88 dB	0.988	35.88 dB	0.988

C. CONTEXT ENCODER/DECODER

The context required for semantic encoding and decoding in the proposed system is shared between the transmitter and the receiver using the key frames of each GOP. Transmission of key frames in their raw format is not feasible due to the excessive bandwidth it will occupy, which offsets the bandwidth saving achieved by semantic compression. Therefore, VVC intra coding, which is the state-of-the-art for compressing video frames, is employed to encode the key frames.

VVC offers a range of intra coding options with different quantization parameter (QP) values where lower values correspond to higher reconstructed frame quality and higher bit rates, and where higher values correspond to lower reconstructed frame quality and lower bit rates. However, unlike in the conventional use of VVC, the key frames of the proposed system do not need to use very low QP, as the semantic decoder primarily relies on the latent vector to reconstruct the frames rather than just on the key frame. Therefore, the effect of varying QP values used to encode key frames for high SI/TI and low SI/TI videos was investigated to identify a suitable QP value that provides a reasonable trade-off between the quality of the reconstructed frame and the bitrate required by the key frames. Table 3 shows the variation of the bitrate required when different QP values are chosen to encode the key frames with VVC intra coding, while maintaining the same QP (22) for encoding the residuals (discussed in Section III-D) for videos with high SI and high TI content.

TABLE 3. Effect of QP value used for encoding key frames for videos with high SI/high TI.

Key frame QP	Key frame rate	PSNR	SSIM	VMAF
22	1,748.9 kbit/s	32.72 dB	0.7866	80.94
27	1,046.5 kbit/s	32.72 dB	0.7853	80.47
32	520.1 kbit/s	32.94 dB	0.7872	79.82
37	210.1 kbit/s	33.04 dB	0.7824	78.08

Table 4 shows the variation of the bitrate under the same conditions for videos with low SI and low TI content.

Based on these observations, using a higher QP value of 37 over a lower value such as 22 only leads to a reduction in rate distortion performance, but provides a significant saving in the bandwidth required in transmitting the key frames. Since the proposed system has a residual coding scheme in place to enable high fidelity reconstructions of the frames at the receiver, a QP of 37 is selected for the VVC intra coding

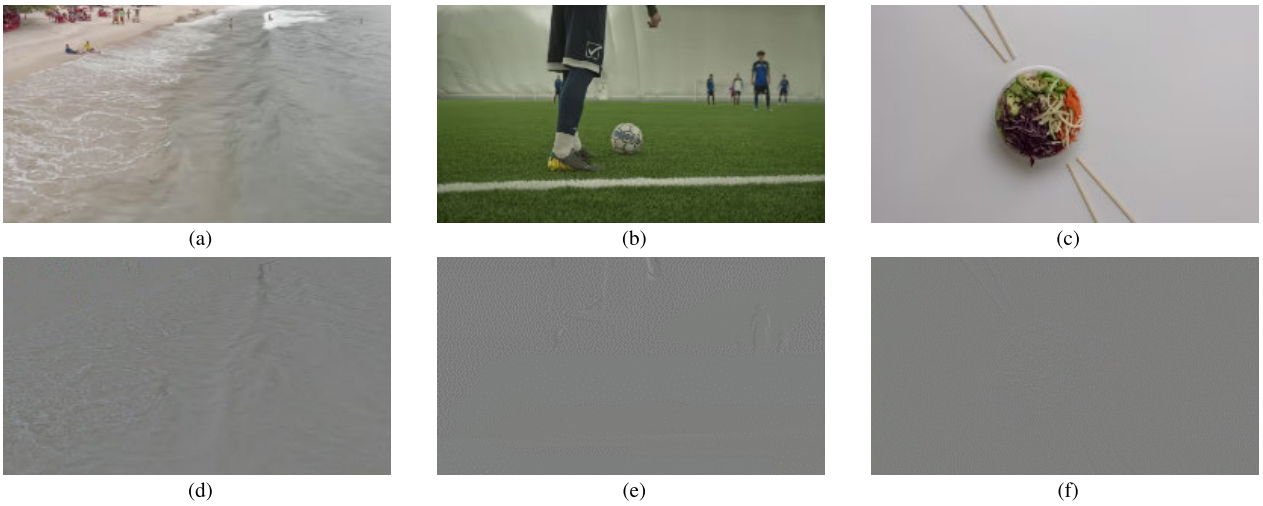


FIGURE 6. Frames from example videos and their corresponding residuals using only semantic and context encoding/decoding. Top row: original frames (a) Video A: medium complexity, (b) Video B: high complexity, (c) Video C: low complexity. Bottom row: corresponding residuals after reconstruction in receiver (d) Video A, (e) Video B, (f) Video C.

TABLE 4. Effect of QP value used for encoding key frames for videos with low SI/low TI.

Key frame QP	Key frame rate	PSNR	SSIM	VMAF
22	947.9 kbit/s	42.71 dB	0.9862	89.80
27	607.2 kbit/s	41.84 dB	0.9819	89.01
32	345.9 kbit/s	40.19 dB	0.9732	86.39
37	187.5 kbit/s	38.17 dB	0.9591	81.02

used to encode key frames such that the bandwidth it occupies is minimized. Section III-D details the implementation of the residual encoding and decoding system used in conjunction with the semantic and context encoding/decoding systems to improve the fidelity of the output video.

D. RESIDUAL ENCODER/DECODER

The semantic encoder/decoder and the context encoder/decoder are capable of reconstructing the video at the receiver with a rate distortion performance that is sufficient for most machine vision applications, such as object identification and classification. However, the fidelity of the reconstructed video has a significant difference, or residual, compared to the original input video, which becomes a major concern for human perception. Fig. 6 shows the residuals of a frame from three example videos with varying levels of complexity, where Fig. 6b is a high complexity video (high SI and/or TI content), Fig. 6a is a medium complexity video (medium SI and medium TI content), and Fig. 6c is a low complexity video (low SI and low TI content). The residuals from difference between the original frames and reconstructed frames of each (Fig. 6d - 6f) are significant, leading to degradation of quality in human vision applications, especially when the video complexity is high (Fig. 6e).

To overcome this challenge, the proposed system implements a residual encoding/decoding mechanism inspired by residual encoding with VVC. As shown in Fig. 2, a semantic decoder and a context decoder with the same structure and configuration as in the receiver are implemented in the transmitter. The semantic decoder is initialized using the Glorot initializer and the same seed as the semantic encoder, and trained using latent vectors and key frames decoded by the context decoder following the same process described in Section III-B. This enables the transmitter to reconstruct a frame nearly identical to that reconstructed at the receiver. Then the reconstructed frame is subtracted from the original frame to obtain a residual with each pixel represented as a signed binary value (9 bits). This is then encoded using VVC with the configuration set to an input and internal bit depth of 10 bits to accommodate the sign bit. A VVC decoder is implemented at the receiver with the same configuration to extract the residual, which is then added to the reconstructed frame from the semantic decoder to produce a high fidelity output.

A key difference between how the residual is encoded in the proposed system and how it is done in conventional video coding systems such as VVC is that the proposed system does not implement any complex motion prediction and motion compensation algorithms in the transmitter. The temporal correlation between frames in a GOP is already exploited in the semantic encoding process, and therefore it only requires minimal prediction capabilities from VVC when encoding the residuals.

However, compared to the bitrate required to transmit the *semantics* and *context* of each frame, the bitrate taken up by the residual is significantly high. Therefore, the GOP size and QP values used for VVC are optimized to minimize the bitrate of the encoded residual after analyzing the variation of the average PSNR and the average SSIM for a range of videos, as shown in Table 5. The observations do not show that a significant gain to be achieved by decreasing the GOP size,

and therefore the highest GOP size of 32, which corresponds to the lowest bit rate of the encoded residuals, is selected.

TABLE 5. Comparison of residual bitrates, PSNR, and SSIM under different GOP sizes for residual encoding using VVC.

GOP Size	Average Bitrate	Average PSNR	Average SSIM
8	43.7 kbit/s	30.22 dB	0.916
16	39.9 kbit/s	30.37 dB	0.916
32	32.6 kbit/s	30.40 dB	0.911

QP values set residual coding using VVC are not fixed and are selected from the range between 8 and 37 depending on the complexity of the video. Given that the latent vector of each frame is constant in size (8×10 bits) and a considerably small value for each frame, this provides the flexibility to adjust the QP dynamically depending on the complexity of the video. Selection of QP values was done manually on a subjective basis for the experiment, but a rule-based configuration can be implemented to select the appropriate QP values for each video or frame to automate the process.

A key enabler to keep the size of the residuals small is to limit the GOPs to within a specific scene of the video, which is implemented using a scene transition detection method as described in Section III-E.

E. SCENE TRANSITION DETECTION

Scene changes in video can be abrupt or gradual, but either of them occurring within a GOP will lead to the input frame to the semantic encoder becoming significantly different from the context (key frame) used by the semantic decoder, which will significantly reduce the fidelity of the semantically decoded frame. Therefore, a scene transition detection algorithm that can effectively determine abrupt and gradual scene changes in uncompressed video input can be applied at the input to the proposed system, as shown in Fig. 2, using the method described in [47]. Scene transitions are used to trigger a new GOP, where the key frame will provide valid context for the semantic encoders, enabling one to keep the fidelity of the semantically decoded image sufficiently high and manage the size of the residuals.

The end-to-end system described in Sections III-A to III-E is implemented and tested using a set of videos with varying complexity and compared with conventional video coding systems in terms of rate distortion performance, of which the process is described in Section III-F.

F. PERFORMANCE EVALUATION

20 video clips of spatial resolution 320×180 pixels and frame rate 20 frames per second (fps), shown in Fig. 7, are used to demonstrate the performance of the system described in Sections III-A to III-E. These videos were selected to represent a range of SI and TI content calculated using [68] and shown in Fig. 8.

To meaningfully summarize the performance of the proposed system and the reference systems, the 20 video clips are classified into three categories based on complexity, as shown in Fig. 8. Videos 1, 3, 4, 5, 7 and 8, which have low SI and low TI, are considered to be low complexity videos for the purpose of analysis. Videos 6, 9, 10, 12, 13, 18 and 20 have medium SI and medium TI and are considered medium complexity videos. Videos that have high SI (Videos 14, 15, 16 and 17) or high TI (Videos 2, 11 and 19) are considered high complexity videos.

The resolution of the video clips selected for testing was restricted by the limitations of computational resources available to train the semantic encoder and semantic decoder in multiple iterations during simulations. However, the resolution of the test videos does not have a significant impact on the rate distortion measures (PSNR, SSIM, and VMAF) used to evaluate the quality of the reconstructed video, since they are based on the input video. This can be verified by considering the variation of the observed PSNR for different spatial resolutions of an example video given in Table 6, which shows only an insignificant change in the metric when the spatial resolution is increased. Therefore, this will not be a limiting factor in scaling up the proposed system for videos of any spatial resolution. Furthermore, the proposed system can also be implemented on blocks partitioned from a high-resolution video using the same principle, which allows it to be scaled on such a basis as well.

TABLE 6. Effect of Spatial Resolution of Video and Observed PSNR for Different Video Coding Systems.

Resolution	Total Pixels	Average PSNR
256×144	110,592	34.32 dB
320×180	172,800	34.33 dB
420×240	306,720	34.33 dB
1280×720	921,600	34.10 dB
1920×1080	2,073,600	34.01 dB

The proposed system is benchmarked against the currently most widely used conventional video coding systems of the H.26X family: VVC, HEVC, and AVC. VVenC 1.8 codec [69] is used as the implementation of VVC and the tools available in ffmpeg [70] are used for the implementation of HEVC and AVC. The parameters used for setting up VVC, HEVC and AVC reference codecs are given in Table 7. The QP values used for residual coding of the proposed system are matched to the QP values used for encoding the test videos with VVC, HEVC and AVC to provide a comparison over a range of quality levels.

Performance comparisons of the proposed system and the three conventional video coding systems are made based on the PSNR, SSIM and VMAF indices and are quantified by calculating the BD rates. PSNR is the relationship between the square of the maximum possible pixel value of an image or video and the mean squared error of the pixel values and is the most common index used to measure the quality of reconstruction of images or videos after they have undergone lossy compression [71]. SSIM is primarily an objective measure of the structural consistency of the decoded video

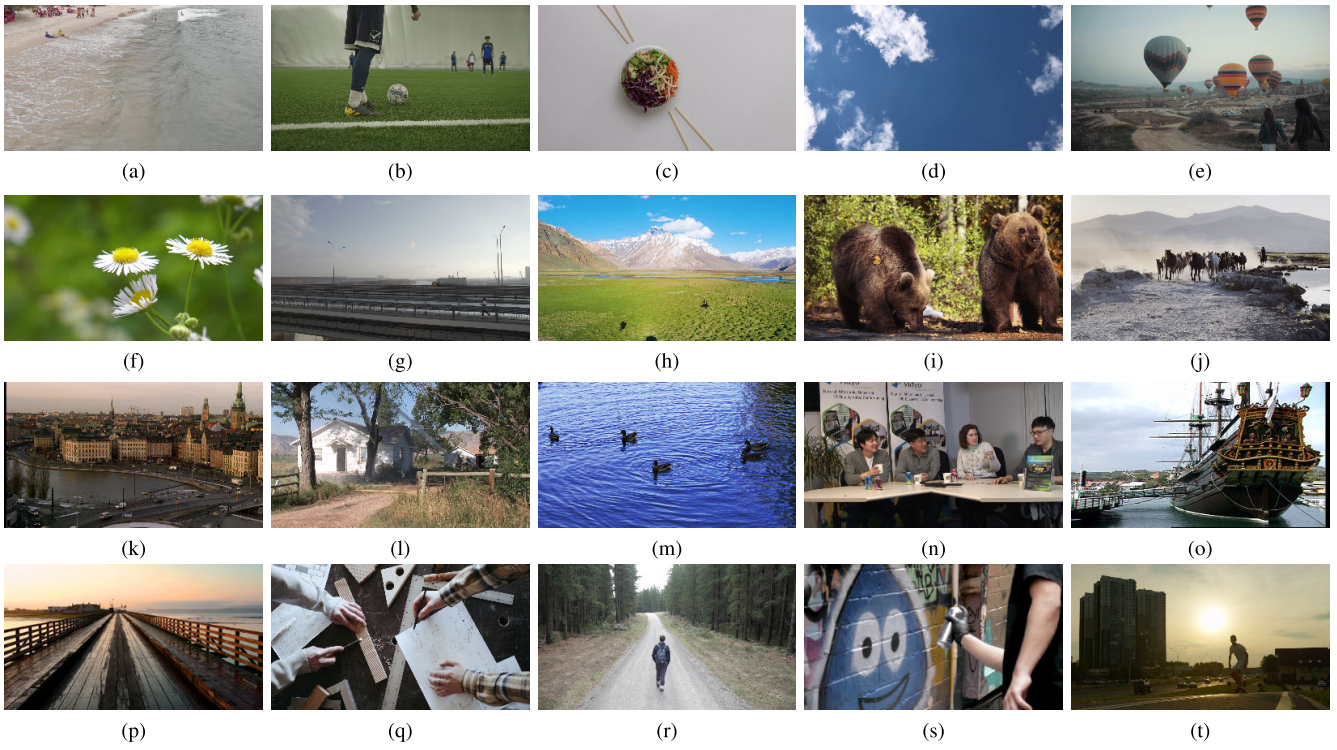


FIGURE 7. Video Clips used for Experiments: (a) Video 1 [48], (b) Video 2 [49], (c) Video 3 [50], (d) Video 4 [51], (e) Video 5 [52], (f) Video 6 [53], (g) Video 7 [54], (h) Video 8 [55], (i) Video 9 [56], (j) Video 10 [57], (k) Video 11 [58], (l) Video 12 [59],(m) Video 13 [60], (n) Video 14 [61], (o) Video 15 [62], (p) Video 16 [63], (q) Video 17 [64], (r) Video 18 [65], (s) Video 19 [66], (t) Video 20 [67].

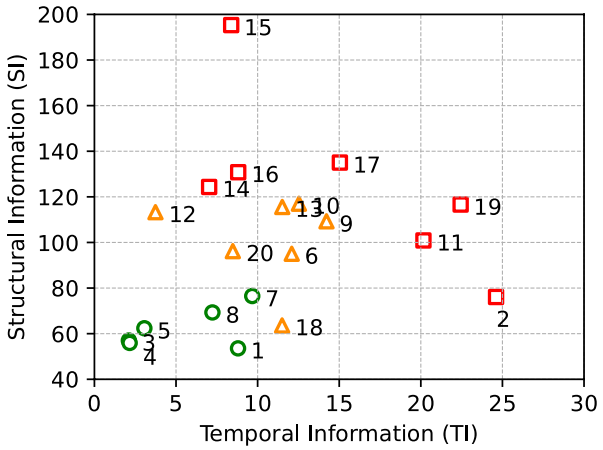


FIGURE 8. Distribution of Structural Information (SI) and Temporal Information (TI) in Test Videos. Categorized as: ∘ - Low Complexity Videos, △ - Medium Complexity Videos, □ - High Complexity Videos.

clip in the spatial domain, and is used here mainly to compare performance in spatial dimension [72]. VMAF, which is an objective measure for both the structural and temporal consistency of the decoded video clip on multiple factors, is used here to mainly compare performance in the temporal dimension [73]. The BD rate [74] is an objective measure that compares the rate distortion performance of a test system with a reference system, with a negative BD rate corresponding to the better performance of the test system compared to the reference system, and a positive BD rate corresponding to a worse performance of the test system compared to the reference system.

TABLE 7. Specifications of VVC, HEVC and AVC used as benchmarks.

Parameter/s	Value
Color space	YUV420
Internal bit depth	8
fps	20
Resolution	320 × 180
GOP	8, 16, 32
QP	22/27/32/37
Preset	Slow
Pixel format	YUV420 4:2:0

The analysis of these metrics with relation to the proposed system and VVC, HEVC, and AVC, and a discussion of their implications are presented in Section IV.

IV. RESULTS AND DISCUSSION

The rate distortion performance of the proposed system is evaluated against VVC, HEVC and AVC for the 20 test video clips, as explained in Section III. The average rate distortion rates (PSNR, SSIM and VMAF) and the corresponding average BD rates are analyzed by grouping the test videos into three groups: low complexity videos, medium complexity videos and high complexity videos based on their SI and TI content. The rate distortion performance of the proposed

system and the three benchmark systems for each individual video for different GOP sizes are shown in Appendix.

A key characteristic of the proposed system is that the latent vector for a frame, regardless of the QP of residual coding and key frames, is constant and small at 2.5 kbit/s since it is capable of effectively extracting the semantics of the frame from a dimensions of 320×180×1.5 to 8×1. Therefore, the majority of the bandwidth consumed by the proposed system consists of the VVC encoded residuals, which are on the order of hundreds of kbit/s.

When GOP size 8 is considered (Fig. 9), rate distortion performance measured by the averages of PSNR, SSIM and VMAF of the proposed system is on par with VVC for low complexity, medium complexity, and high complexity videos. It also shows a significant coding gain over HEVC and AVC, especially for QP values of 27 and QP 22 when the GOP size is set to 8. These improvements can be quantified using BD rates, as shown in Figs. 9j, 9k, and 9l. The proposed system is capable of outperforming VVC in terms of PSNR, SSIM, and VMAF under test conditions when the video complexity is low or medium when the GOP size is 8. However, VVC has marginally better rate distortion performance in PSNR and SSIM when video complexity is high, although it falls behind the proposed system in terms of VMAF. These results demonstrate that the proposed system is capable of significantly improving the rate distortion performance compared to HEVC and AVC, while matching that of VVC, when smaller GOP sizes, such as 8, are considered. Table 8 summarizes the BD rates for each rate distortion metric with reference to AVC, HEVC, and VVC.

TABLE 8. Average BD Rates for Proposed System with Reference to AVC, HEVC and VVC when GOP Size = 8.

Metric	Complexity	vs. AVC	vs. HEVC	vs. VVC
PSNR	Low	-32.28	-21.37	-17.49
PSNR	Medium	-36.27	-19.16	-13.93
PSNR	High	-31.62	-19.69	-12.74
SSIM	Low	-37.09	-20.47	-16.62
SSIM	Medium	-46.56	-28.65	-20.91
SSIM	High	-41.17	-25.45	-15.67
VMAF	Low	-32.04	-23.98	-21.2
VMAF	Medium	-33.98	-19.07	-8.88
VMAF	High	-30.67	-21.41	-12.01

When GOP size 16 is considered (Fig. 10), similar to GOP size 8, the rate distortion performance of the proposed system in terms of PSNR, SSIM and VMAF for low, medium, and high complexity videos demonstrate clear improvements over HEVC and AVC. The improvements achieved compared to VVC are generally on par when considering PSNR and SSIM metrics. However, VVC slightly outperforms the proposed system in terms of average VMAF for low and medium complexity videos when GOP size of 16 is considered. Comparison of the performance of rate distortion on PSNR, SSIM, and VMAF is clearly observable with the BD rates shown in Figs. 10j, 10k, and 10l. Table 9 summarizes the BD rates for each rate distortion metric with reference to AVC, HEVC, and VVC.

TABLE 9. Average BD Rates for Proposed System with Reference to AVC, HEVC and VVC when GOP Size = 16.

Metric	Complexity	vs. AVC	vs. HEVC	vs. VVC
PSNR	Low	-53.29	-48.83	-15
PSNR	Medium	-46.53	-36.75	0.84
PSNR	High	-55.71	-50.88	-20.85
SSIM	Low	-58.73	-50.46	-11.16
SSIM	Medium	-57.38	-47.08	-12.38
SSIM	High	-62.5	-55.33	-21.58
VMAF	Low	-53.95	-51.66	-15.79
VMAF	Medium	-44.35	-36.75	5.4
VMAF	High	-56.95	-54.84	-18.83

TABLE 10. Average BD Rates for Proposed System with Reference to AVC, HEVC and VVC when GOP Size = 32.

Metric	Complexity	vs. AVC	vs. HEVC	vs. VVC
PSNR	Low	-57.14	-54.19	-18.33
PSNR	Medium	-48.2	-41.7	2.68
PSNR	High	-58.57	-56.26	-19.46
SSIM	Low	-64.81	-58.34	-19.78
SSIM	Medium	-56.97	-47.6	-5.88
SSIM	High	-64.05	-59.36	-18.94
VMAF	Low	-56.06	-54.78	-16.86
VMAF	Medium	-50.14	-44.66	0.77
VMAF	High	-61.08	-61.17	-21.56

The proposed system again demonstrates a rate distortion performance comparable to that of VVC and significantly better than that of HEVC and AVC when considering a GOP size of 32 (Fig. 11). It has significantly better rate distortion performance compared to HEVC and VVC across all three categories of video complexity for a GOP size of 32. Compared to VVC, the proposed system demonstrates better PSNR and SSIM performance when video complexity is low or high, but VVC outperforms it when video complexity is medium. When VMAF is considered, the proposed system outperforms VVC when the complexity of the video is medium or high, but VVC shows better performance when the complexity of the video is low. BD rates, as shown in Figs. 11j, 11k, and 11l, confirm these observations for a GOP size 32. Table 10 summarizes the BD rates for each rate distortion metric with reference to AVC, HEVC, and VVC.

Overall, the proposed system has slightly lower rate distortion performance compared to VVC, especially when SSIM and VMAF are considered, when the GOP sizes are 16 and 32, and when the video complexity is medium (corresponding to medium SI and medium TI). When the video complexity is high (corresponding to high SI or high TI) or low (corresponding to low SI and low TI), the proposed system outperforms VVC under the given test conditions. The reason for VVC outperforming the proposed system for videos with medium complexity is that while semantic based frame prediction is capable of capturing inter frame

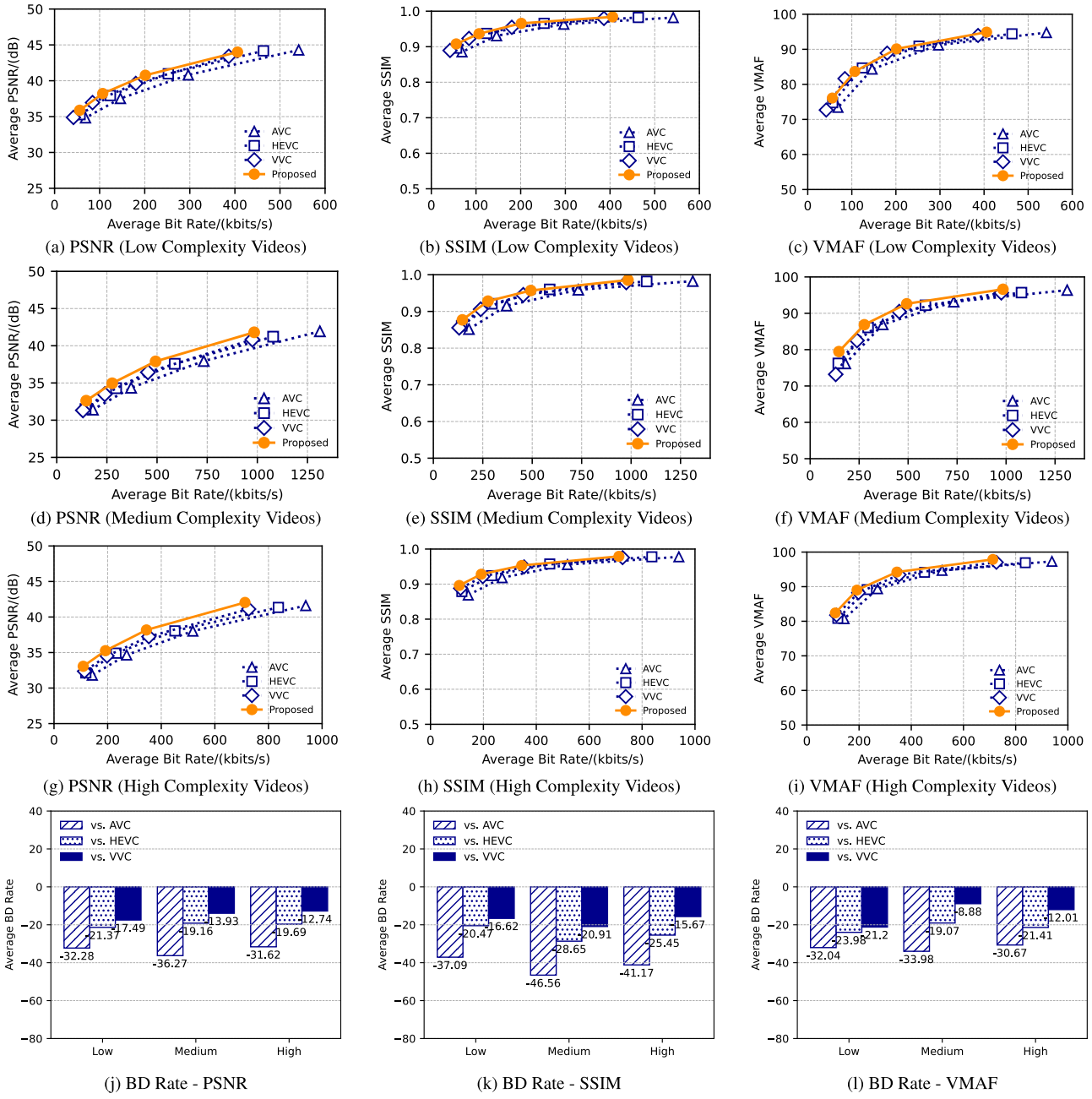


FIGURE 9. Performance Comparison of Proposed System with VVC, HEVC, and AVC measured by PSNR, SSIM and VMAF and corresponding BD rates when GOP size is 8. First Row (a, b, c): Averages for Low Complexity Videos. Second Row (d, e, f): Averages for Medium Complexity Videos. Third Row (g, h, i): Averages for High Complexity Videos. Fourth row: (j) PSNR BD rate, (k) SSIM BD rate, (l) VMAF BD rate.

information in the temporal domain through the GOP wise training process, it lacks the complexity and capabilities in VVC, which has evolved over decades, to perform predictions using motion estimation and motion compensation. However, when the complexity of the video is high, the nonlinearity of the transformation implemented in the proposed system gives the advantage of requiring a lower bit rate to transmit the video compared to VVC.

VVC predicts frames based on motion vectors of the different blocks that are used to derive predicted frames (P frames) and bidirectional frames (B frames), which have evolved from the techniques used in AVC and HEVC [75]. Calculation of these motion vectors and prediction and encoding of motion based on a range of cost functions contribute significantly to computational complexity in VVC. However, since most of the methods that implement motion estimation, such as those discussed in [76], are linear transformations optimized for medium complexity videos, which are the most common types, it invariably leads to an increase of the bit rate required when encountered with

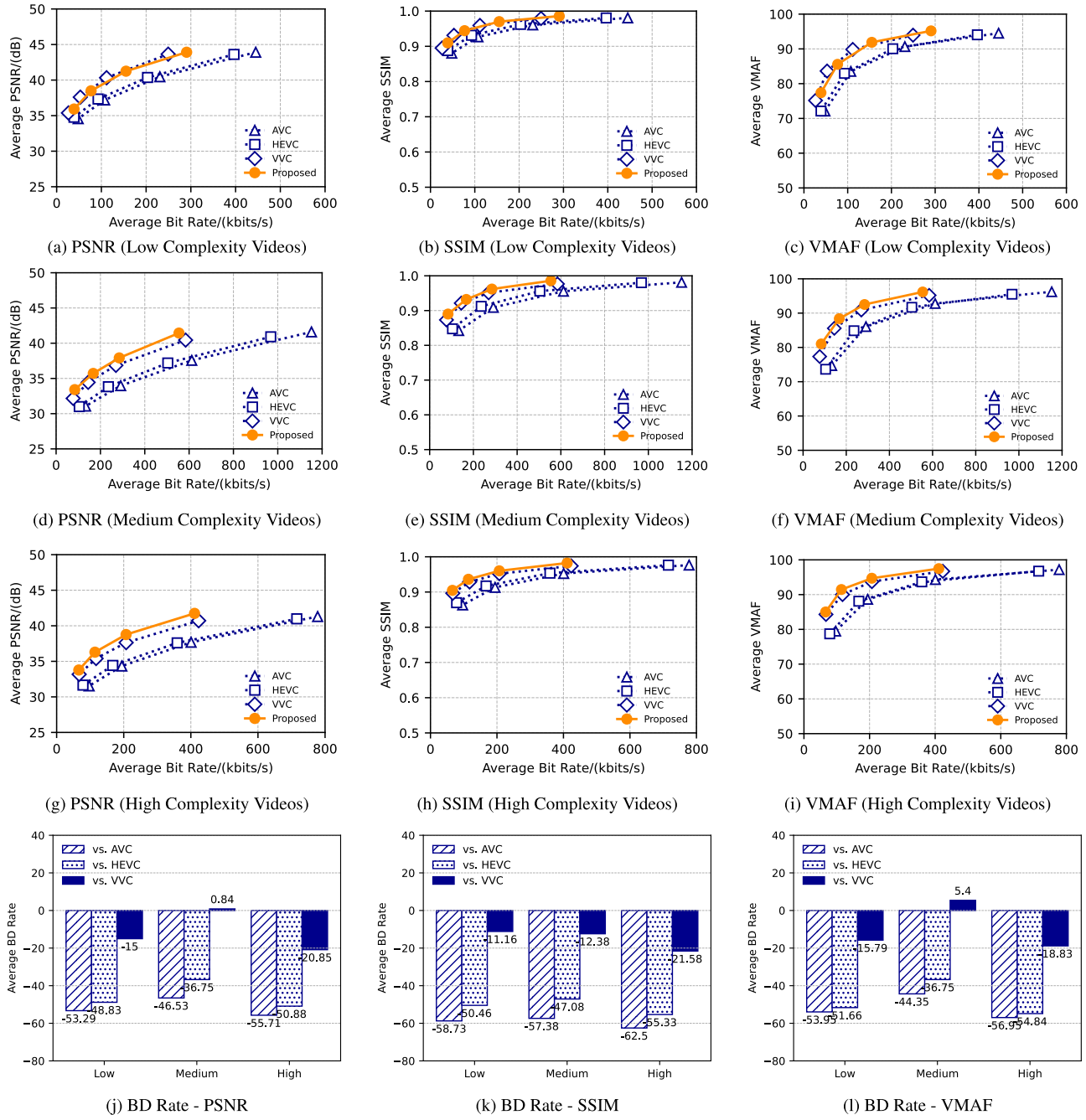


FIGURE 10. Performance Comparison of Proposed System with VVC, HEVC, and AVC measured by PSNR, SSIM and VMAF and corresponding BD rates when GOP size is 16. First Row (a, b, c): Averages for Low Complexity Videos. Second Row (d, e, f): Averages for Medium Complexity Videos. Third Row (g, h, i): Averages for High Complexity Videos. Fourth row: (j) PSNR BD rate, (k) SSIM BD rate, (l) VMAF BD rate.

a video with high TI and high SI. However, the nonlinear nature of the transformation implemented by the semantic encoding process keeps the bitrate of the latent vectors constant regardless of the complexity of the frame. The trade-off is with the size of the residual, which increases in size compared to VVC when the video complexity increases from low to medium. But when the video complexity becomes high, the non linear encoding of the semantic encoder is able to better predict the frames, leading to lesser residuals which improve the overall rate distortion performance.

In terms of encoder complexity, the main burden of the proposed system is the training process of the AE encoder layers and the AE decoder layers, which must be carried out over multiple epochs to achieve convergence. The AEs used are optimized by minimizing the number of layers and implementing uniform initialization across each to ensure that each AE is initialized with the same

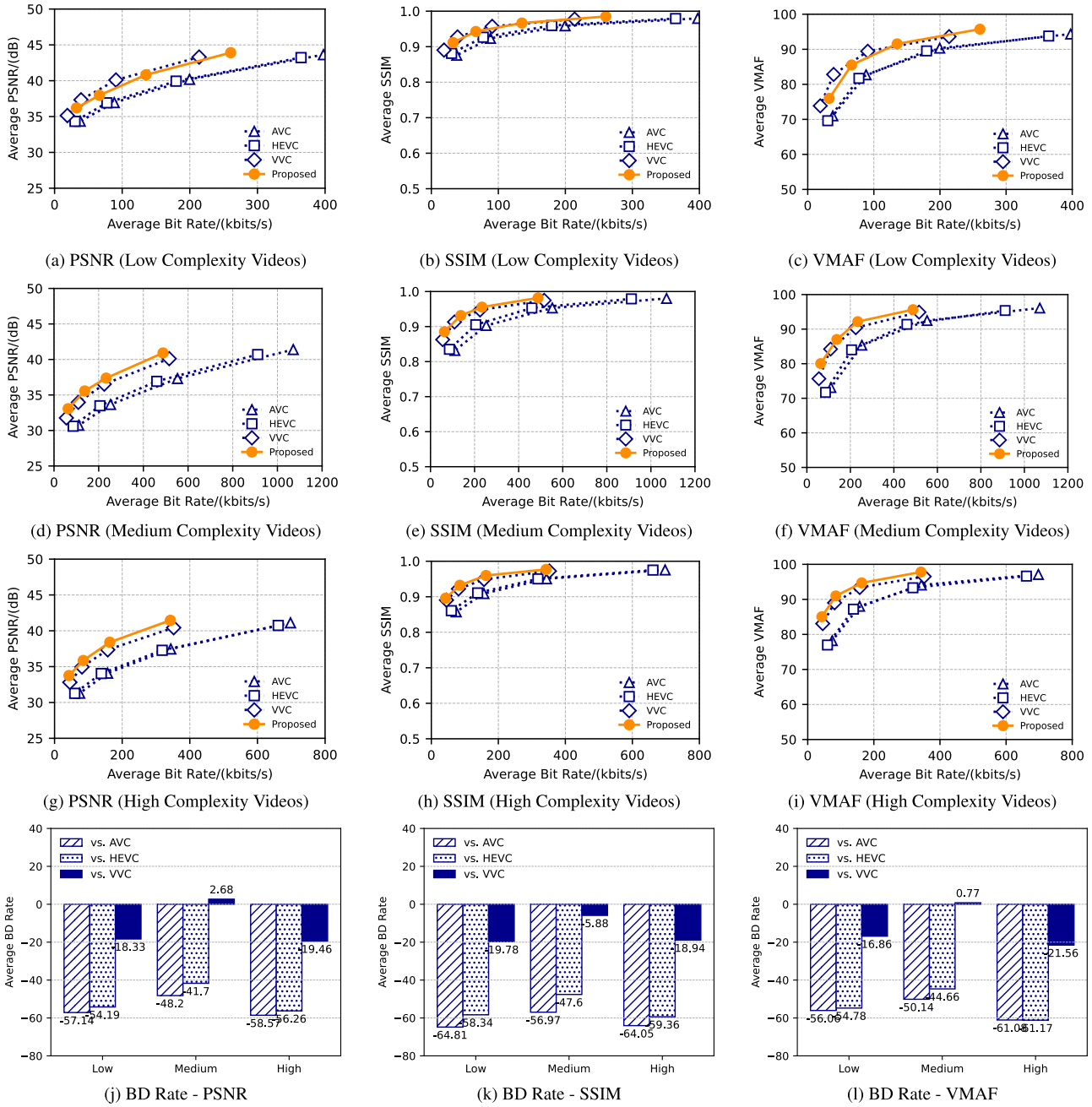


FIGURE 11. Performance Comparison of Proposed System with VVC, HEVC, and AVC measured by PSNR, SSIM and VMAF and corresponding BD rates when GOP size is 32. First Row (a, b, c): Averages for Low Complexity Videos. Second Row (d, e, f): Averages for Medium Complexity Videos. Third Row (g, h, i): Averages for High Complexity Videos. Fourth row: (j) PSNR BD rate, (k) SSIM BD rate, (l) VMAF BD rate.

weights, minimizing the MSE at the output. While the fundamentally different architectures used by the proposed system and conventional video coding systems make a direct comparison of complexity meaningless, the computational complexity of the proposed system can be analyzed in terms of floating point operations (FLOPs), which is an estimate of the computational cost, and multiply-accumulate operations (MACs), which is an estimate of the memory usage. For videos with spatial resolution 320 × 180 and YUV420 pixel format (corresponding to 1.5 color channels) used to test the proposed system, FLOPs and MACs are calculated with reference to the AE architecture shown in Fig. 3.

In the semantic encoder, the first flatten layer does not involve any computations that contribute to FLOPs and just converts the tensor with dimensions 320 × 180 × 1.5 to a vector with dimensions 86, 400 × 1. The first dense layer fully connects the input layer to a hidden layer of dimensions 1024 × 1 corresponding to 176,947,200 FLOPs, which subsequently fully connects to the bottleneck layer of dimensions 8 × 1 corresponding to 16,384 FLOPs. The AE

decoding layers follow a similar setup in reverse, and the reshape layer at the end does not contribute to any FLOPs. Therefore, the total FLOPs for the AE based encoder and decoder approximate 353,927,168. The same AE architecture (shown in Fig. 3) yields approximately 176,963,584 MAC operations.

Although an identical metric is not available to objectively compare the computational complexity of VVC (as implemented in VVenC), a subjective comparison shows that VVenC, as used in the proposed system, is required to perform considerably less motion estimation due to the temporal redundancies already being captured in the semantic encoding process. Since motion estimation is the most resource intensive module in VVC [77], the ability of the proposed system to significantly minimize the motion estimation required should contribute to significantly less computational complexity taken up for residual coding in the proposed system.

A rough estimate of the possible reduction in computational complexity with the proposed system can be made considering the time taken to encode a video using VVenC with and without the use of the semantic encoder. On the same hardware, the average time taken to encode the residual of the semantic encoder with VVenC is 0.765 s/frame, while the time taken to encode the same video with VVenC without using the semantic encoder is 2.026 s/frame. Therefore, it can be assumed that the semantic encoding process enables a significant reduction in computational complexity, although it is necessary to train the semantic encoder/decoder for each GOP and scene.

The proposed system is the first attempt to propose a semantic communication based hybrid video codec, where the semantic decoder is remotely trained using latent vectors, key frames, and residual information that makes it unnecessary to transmit the trained decoder to the receiver or implement a feedback channel from the receiver to the transmitter [78] to allow the use of a trained decoder in the receiver. The dynamic retraining of the AE based neural network for each GOP also enables the proposed system to transmit video clips of increased complexity involving multiple scene changes, which will not be possible with a static trained neural network based decoder. However, there are still multiple open research areas that must be covered to realize the proposed system, which can consistently outperform conventional video coding systems.

The proposed system employs a semantic decoder that attempts to reconstruct predicted frames based on a decoder network trained with only the GOP key frame. Although this approach reduces the amount of context information that needs to be exchanged between the transmitter and receiver in the form of key frames, it also increases the entropy of the residuals that are created between the predicted and original frames. This leads to more bandwidth being required to encode the residual, which diminishes the coding gain that can be achieved from using a semantic encoder to encode the predicted frames instead of a conventional frame encoder. Therefore, further improvements in frame prediction, as well as scene detection, can be incorporated into the proposed system to improve the prediction accuracy, and thereby reduce the entropy of the residuals.

Another approach to improve the frame prediction capability using the semantic decoder is to implement hierarchical frame prediction, where rather than sequentially reconstructing the predicted frames using the key frames, a hierarchical approach (e.g. for a GOP of 8, use frames 1 and 9 to predict frame 5, frame 1 and frame 5 to predict frame 3, frame 1 and frame 3 to predict frame 2 etc.). Adopting such a hierarchical prediction approach enables the predicted frames to be better approximations of the original and leads to decreasing the entropy of the residuals from the predicted frames.

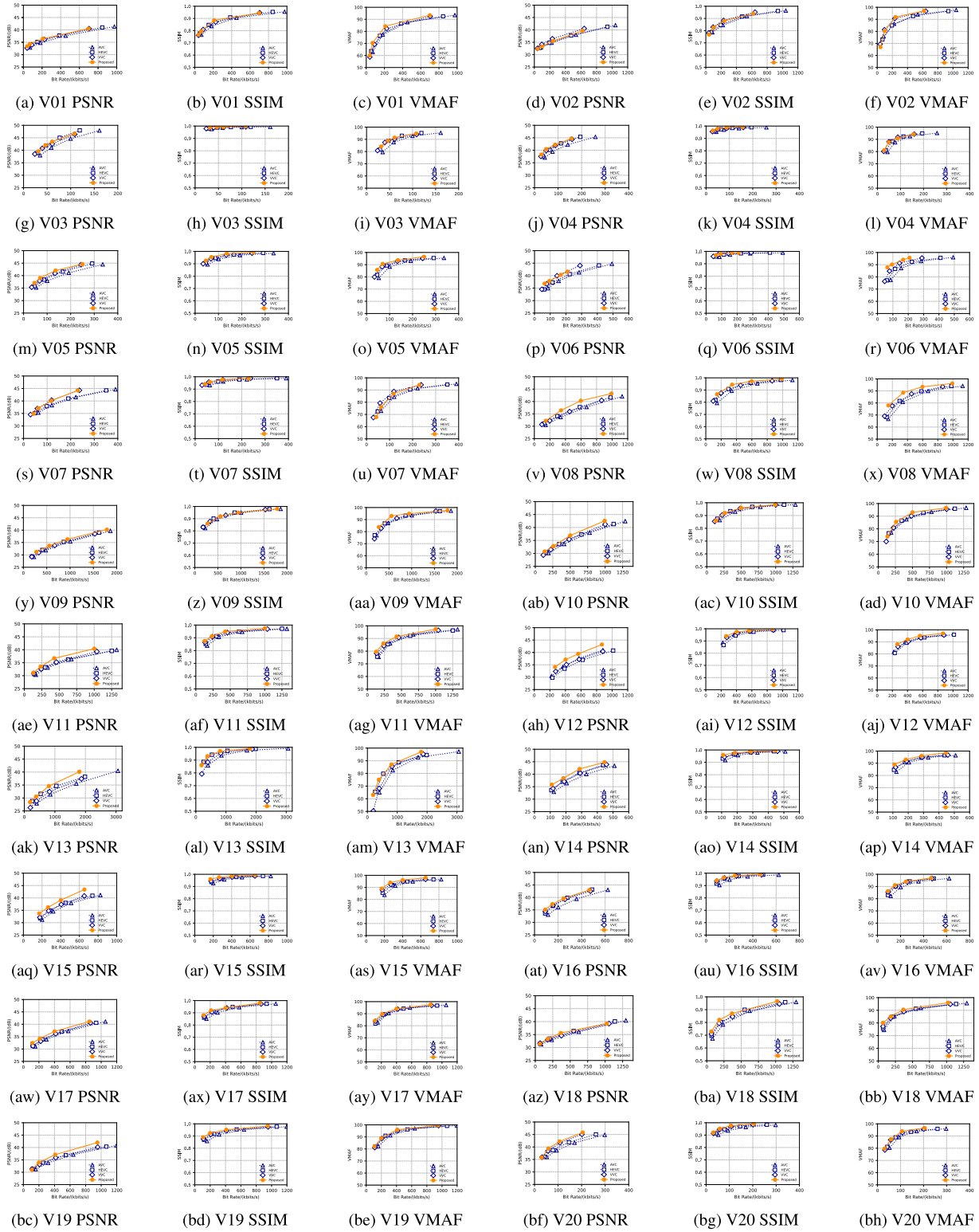
In terms of complexity, the proposed system has the additional task of training neural networks compared to conventional video coding systems, which will add computational cost, but will also provide the benefits of a lower bitrate and a more resilient base layer to combat channel noise. This complexity can be traded off if the proposed video coding system is able to encode the residuals, avoiding the use of VVC encoding, as discussed above. Furthermore, future research on preprocessing, optimizing the autoencoder design, and optimizing the training strategy making use of inherent redundancies and correlations present in video can drastically reduce the computational cost, while adapting optimized hardware for neural network deployment can significantly lower the latency of the proposed system.

Another key area for future research, especially for applications intended for human viewing, is the implementation of an effective residual coding mechanism to transmit the residuals of the image more efficiently between the transmitter and the receiver. Currently, it is based on VVC, which is not optimized for residuals resulting from semantic encoding and decoding. Resolving this problem with an innovative approach will not only enable the proposed video coding system to drastically reduce the bandwidth it needs to operate effectively but can also provide a means to improve the coding performance of conventional video coding systems as well, since residual coding uses up most of the data rate required for both systems.

In summary, the proposed system has shown that it can perform better than VVC, HEVC, and AVC for most video coding scenarios involving low and high complexity videos under test conditions, and it can closely follow the performance of VVC rate distortion for medium complexity videos. Thus, the concept, with the refinements discussed, has enormous potential to be adapted to a wide range of next-generation video coding paradigms.

V. CONCLUSION

We propose a novel hybrid video codec based on semantic communication and VVC, which incorporates concepts of semantic communications to create a latent vector of video frames augmented by residual coding capabilities of VVC,



(a) V01 PSNR (b) V01 SSIM (c) V01 VMAF (d) V02 PSNR (e) V02 SSIM (f) V02 VMAF

(g) V03 PSNR (h) V03 SSIM (i) V03 VMAF (j) V04 PSNR (k) V04 SSIM (l) V04 VMAF

(m) V05 PSNR (n) V05 SSIM (o) V05 VMAF (p) V06 PSNR (q) V06 SSIM (r) V06 VMAF

(s) V07 PSNR (t) V07 SSIM (u) V07 VMAF (v) V08 PSNR (w) V08 SSIM (x) V08 VMAF

(y) V09 PSNR (z) V09 SSIM (aa) V09 VMAF (ab) V10 PSNR (ac) V10 SSIM (ad) V10 VMAF

(ae) V11 PSNR (af) V11 SSIM (ag) V11 VMAF (ah) V12 PSNR (ai) V12 SSIM (aj) V12 VMAF

(ak) V13 PSNR (al) V13 SSIM (am) V13 VMAF (an) V14 PSNR (ao) V14 SSIM (ap) V14 VMAF

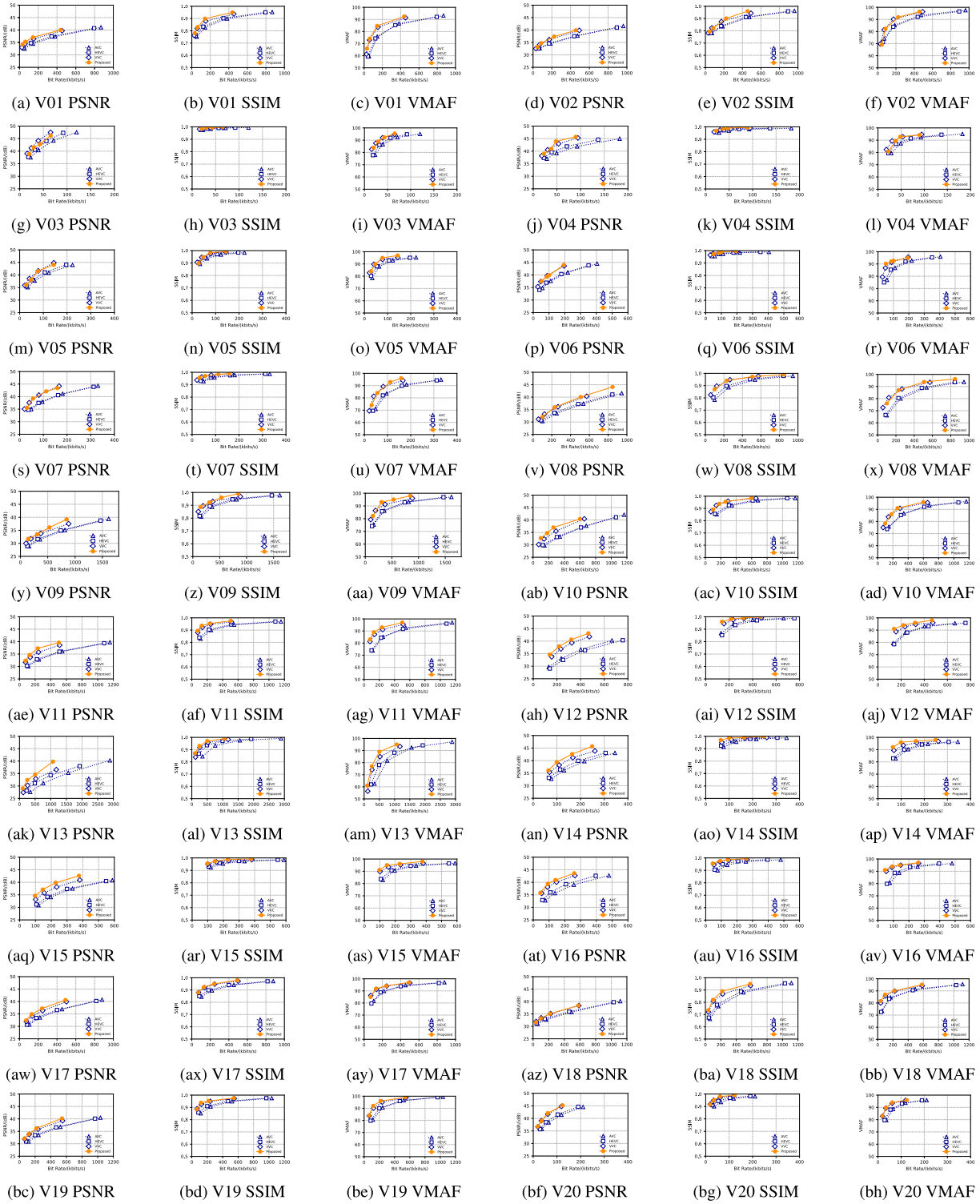
(aq) V15 PSNR (ar) V15 SSIM (as) V15 VMAF (at) V16 PSNR (au) V16 SSIM (av) V16 VMAF

(aw) V17 PSNR (ax) V17 SSIM (ay) V17 VMAF (az) V18 PSNR (ba) V18 SSIM (bb) V18 VMAF

(bc) V19 PSNR (bd) V19 SSIM (be) V19 VMAF (bf) V20 PSNR (bg) V20 SSIM (bh) V20 VMAF

FIGURE 12. Rate distortion performance (PSNR, SSIM and VMAF) for each test video clip when GOP size is 8.

which can achieve significant coding gains over VVC, HEVC and AVC for videos with low complexity (low SI and low TI) and high complexity (high SI or high TI). It also introduces a novel approach to using DNN based video encoders in real-world scenarios, where the receiver-side decoder can be trained remotely in coordination with the transmitter-side



(a) V01 PSNR (b) V01 SSIM (c) V01 VMAF (d) V02 PSNR (e) V02 SSIM (f) V02 VMAF

(g) V03 PSNR (h) V03 SSIM (i) V03 VMAF (j) V04 PSNR (k) V04 SSIM (l) V04 VMAF

(m) V05 PSNR (n) V05 SSIM (o) V05 VMAF (p) V06 PSNR (q) V06 SSIM (r) V06 VMAF

(s) V07 PSNR (t) V07 SSIM (u) V07 VMAF (v) V08 PSNR (w) V08 SSIM (x) V08 VMAF

(y) V09 PSNR (z) V09 SSIM (aa) V09 VMAF (ab) V10 PSNR (ac) V10 SSIM (ad) V10 VMAF

(ae) V11 PSNR (af) V11 SSIM (ag) V11 VMAF (ah) V12 PSNR (ai) V12 SSIM (aj) V12 VMAF

(ak) V13 PSNR (al) V13 SSIM (am) V13 VMAF (an) V14 PSNR (ao) V14 SSIM (ap) V14 VMAF

(aq) V15 PSNR (ar) V15 SSIM (as) V15 VMAF (at) V16 PSNR (au) V16 SSIM (av) V16 VMAF

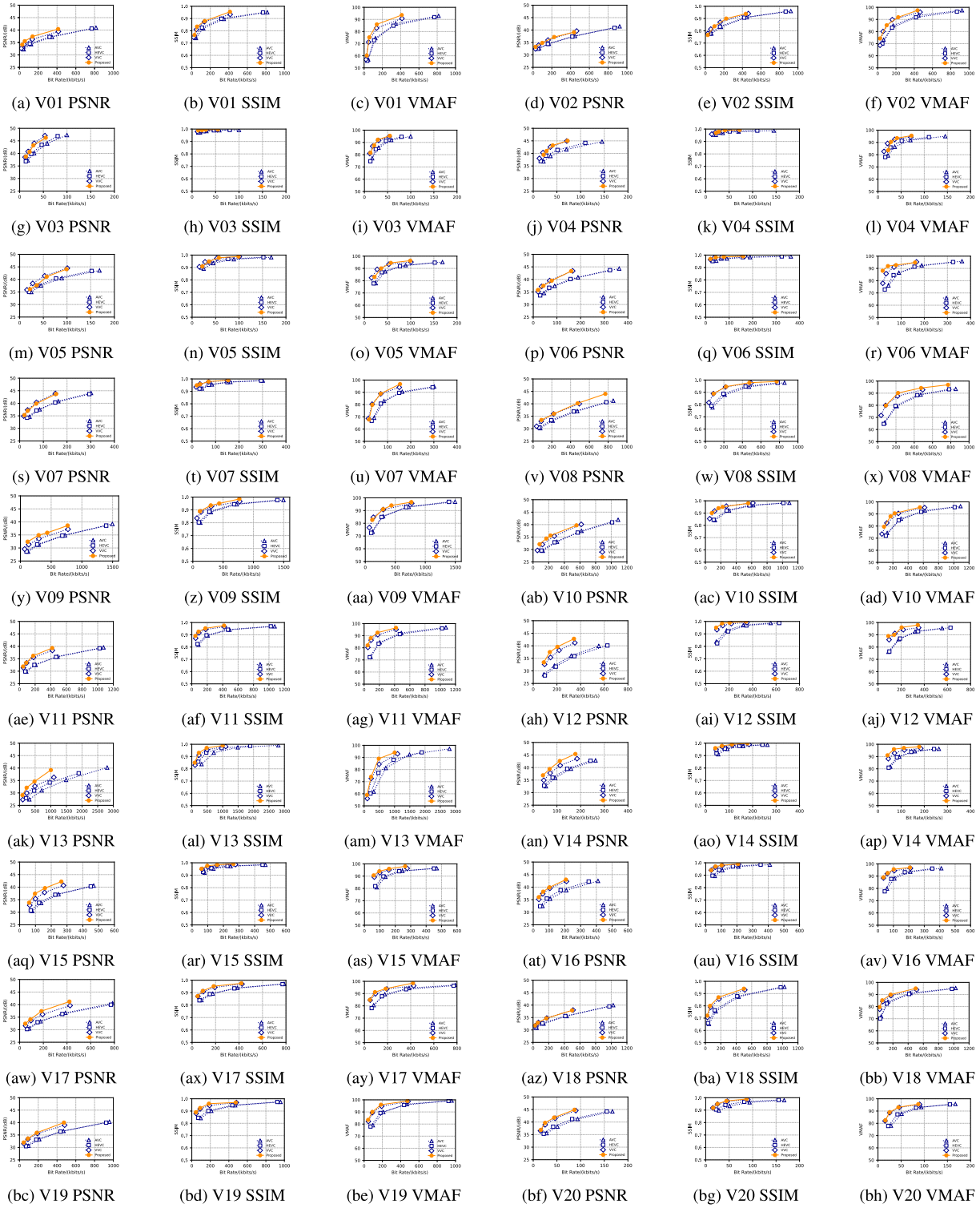
(aw) V17 PSNR (ax) V17 SSIM (ay) V17 VMAF (az) V18 PSNR (ba) V18 SSIM (bb) V18 VMAF

(bc) V19 PSNR (bd) V19 SSIM (be) V19 VMAF (bf) V20 PSNR (bg) V20 SSIM (bh) V20 VMAF

FIGURE 13. Rate distortion performance (PSNR, SSIM and VMAF) for each test video clip when GOP size is 16.

encoder, removing the burden of having to compress and transmit the DNN parameters to the receiver, which will negate any coding gain and noise resilience obtained by using semantic communications concepts.

When compared to conventional video codecs, VVC, HEVC, and AVC, the proposed system consistently outperforms HEVC and AVC with significant coding gains quantified by BD rates in terms of PSNR, SSIM, and VMAF



(a) V01 PSNR (b) V01 SSIM (c) V01 VMAF (d) V02 PSNR (e) V02 SSIM (f) V02 VMAF

(g) V03 PSNR (h) V03 SSIM (i) V03 VMAF (j) V04 PSNR (k) V04 SSIM (l) V04 VMAF

(m) V05 PSNR (n) V05 SSIM (o) V05 VMAF (p) V06 PSNR (q) V06 SSIM (r) V06 VMAF

(s) V07 PSNR (t) V07 SSIM (u) V07 VMAF (v) V08 PSNR (w) V08 SSIM (x) V08 VMAF

(y) V09 PSNR (z) V09 SSIM (aa) V09 VMAF (ab) V10 PSNR (ac) V10 SSIM (ad) V10 VMAF

(ae) V11 PSNR (af) V11 SSIM (ag) V11 VMAF (ah) V12 PSNR (ai) V12 SSIM (aj) V12 VMAF

(ak) V13 PSNR (al) V13 SSIM (am) V13 VMAF (an) V14 PSNR (ao) V14 SSIM (ap) V14 VMAF

(aq) V15 PSNR (ar) V15 SSIM (as) V15 VMAF (at) V16 PSNR (au) V16 SSIM (av) V16 VMAF

(aw) V17 PSNR (ax) V17 SSIM (ay) V17 VMAF (az) V18 PSNR (ba) V18 SSIM (bb) V18 VMAF

(bc) V19 PSNR (bd) V19 SSIM (be) V19 VMAF (bf) V20 PSNR (bg) V20 SSIM (bh) V20 VMAF

FIGURE 14. **Rate distortion performance (PSNR, SSIM and VMAF) for each test video clip when GOP size is 32.**

for all videos tested. It also outperforms VVC when the video complexity is low or high, while falling slightly behind when the video complexity is medium. The key challenges identified in moving forward with the proposed system is optimizing the entropy of the residual through better frame prediction and optimizing the residual coding mechanism to make it optimized for semantic communications and independent of VVC. Overcoming these challenges will

allow the proposed system to consistently outperform VVC, the state-of-the-art conventional video coding system, in a wide range of video applications.

APPENDIX
INDIVIDUAL RATE DISTORTION PERFORMANCE FOR TEST VIDEOS FOR GOP SIZE 8, 16 AND 32

The individual rate distortion performance for each test video under GOP sizes of 8, 16 and 32 are shown in Fig. 12, Fig. 13 and Fig. 14 respectively. For each video clip, the performances of PSNR, SSIM and VMAF of the proposed system (orange with the symbol ○) as well as those of VVC (blue with the symbol ◊), HEVC (blue with the symbol □) and AVC (blue with the symbol △) are shown. Individual video clips are identified by the prefix V and two digits to represent the video number, as referred to in Fig. 7.

The figures in the main text show a summarized version of these results by presenting the averages of PSNR, SSIM and VMAF when the 20 test videos are grouped in terms of complexity.

ACKNOWLEDGMENT

The authors would like to thank the editors and the anonymous reviewers from IEEE Access for their constructive and insightful comments and suggestions which helped in substantially improving the presentation of this article.

REFERENCES

[1] Ericsson. (2023). *Video Traffic Update*. Stockholm. [Online]. Available: https://www.ericsson.com/en/reports-and-papers/mobility-report/dataforecasts/traffic-by-application
[2] Vcodex. (2010). *Historical Timeline of Video Coding Standards and Formats*. Accessed: Aug. 2, 2023. [Online]. Available: https://www.vcodex.com/historical-timeline-of-video-coding-standards-and-formats/
[3] D. García-Lucas, G. Cebrián-Márquez, and P. Cuenca, "Rate-distortion/complexity analysis of HEVC, VVC and AV1 video codecs," *Multimedia Tools Appl.*, vol. 79, nos. 39–40, pp. 29621–29638, Oct. 2020.
[4] A. Mercat, A. Mäkinen, J. Sainio, A. Lemmetti, M. Viitanen, and J. Vanne, "Comparative rate-distortion-complexity analysis of VVC and HEVC video codecs," *IEEE Access*, vol. 9, pp. 67813–67828, 2021.
[5] D. Liu, Y. Li, J. Lin, H. Li, and F. Wu, "Deep learning-based video coding: A review and a case study," *ACM Comput. Surv.*, vol. 53, no. 1, pp. 1–35, Feb. 2020, doi: 10.1145/3368405.
[6] C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication*. Urbana, IL, USA: Univ. Illinois Press, 1949.
[7] Z. Qin, X. Tao, J. Lu, W. Tong, and G. Ye Li, "Semantic communications: Principles and challenges," 2021, *arXiv:2201.01389*.
[8] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, Jul. 2006, doi: 10.1126/science.1127647.
[9] A. A. Alatan, L. Onural, M. Wollborn, R. Mech, E. Tuncel, and T. Sikora, "Image sequence analysis for emerging interactive multimedia services—The European COST 211 framework," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 8, no. 7, pp. 802–813, Nov. 1998.
[10] L. Hanzo, P. Cherriman, and J. Streit, *Video Compression and Communications: From Basics to H.261, H.263, H.264, MPEG4 for DVB and HSDPA-Style Adaptive Turbo-Transceivers*. Chichester, U.K.: Wiley, 2007.
[11] B. Bross, K. Andersson, M. Bläser, V. Drugeon, S.-H. Kim, J. Lainema, J. Li, S. Liu, J.-R. Ohm, G. J. Sullivan, and R. Yu, "General video coding technology in responses to the joint call for proposals on video compression with capability beyond HEVC," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 5, pp. 1226–1240, May 2020.
[12] B. Bross, J. Chen, J.-R. Ohm, G. J. Sullivan, and Y.-K. Wang, "Developments in international video coding standardization after AVC, with an overview of versatile video coding (VVC)," *Proc. IEEE*, vol. 109, no. 9, pp. 1463–1493, Sep. 2021.
[13] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1649–1668, Dec. 2012.
[14] M. T. Pourazad, C. Doutre, M. Azimi, and P. Nasiopoulos, "HEVC: The new gold standard for video compression: How does HEVC compare with H.264/AVC?" *IEEE Consum. Electron. Mag.*, vol. 1, no. 3, pp. 36–46, Jul. 2012.
[15] L. Mou, X. Chen, T. Huang, and W. Gao, "Overview of IEEE 1857.3: Systems of advanced audio and video coding," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Jul. 2014, pp. 1–4.
[16] P. K. Aeluri, V. Bojan, S. Richie, and A. Weeks, "Objective quality analysis of MPEG-1, MPEG-2 & windows media video," in *Proc. 6th IEEE Southwest Symp. Image Anal. Interpretation*, Sep. 2004, pp. 221–225.
[17] Xiph.org Foundation. (2008). *Theora Video Compression*. Accessed: Jul. 19, 2023. [Online]. Available: https://www.theora.org
[18] J. Bankoski, P. Wilkins, and Y. Xu, "Technical overview of VP8, an open source video codec for the web," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2011, pp. 1–6.
[19] D. Mukherjee, J. Han, J. Bankoski, R. Bultje, A. Grange, J. Koleszar, P. Wilkins, and Y. Xu, "A technical overview of VP9—The latest open-source video codec," *SMPTE Motion Imag. J.*, vol. 124, no. 1, pp. 44–54, Jan. 2015.
[20] Alliance for Open Media. (2023). Accessed: Aug. 2, 2023. [Online]. Available: https://aomedia.org/av1/
[21] G. J. Sullivan and T. Wiegand, "Video compression—From concepts to the H.264/AVC standard," *Proc. IEEE*, vol. 93, no. 1, pp. 18–31, Jan. 2005.
[22] S. Ma, X. Zhang, C. Jia, Z. Zhao, S. Wang, and S. Wang, "Image and video compression with neural networks: A review," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 6, pp. 1683–1698, Jun. 2020.
[23] G. Lu, W. Ouyang, D. Xu, X. Zhang, C. Cai, and Z. Gao, "DVC: An end-to-end deep video compression framework," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10998–11007.
[24] D. Alexandre, H.-M. Hang, W.-H. Peng, and M. Domanski, "Deep video compression for interframe coding," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2021, pp. 2124–2128.
[25] A. Mercat, M. Viitanen, and J. Vanne, "Uvg dataset: 50/120fps 4k sequences for video codec analysis and development," in *Proc. 11th ACM Multimedia Syst. Conf.* New York, NY, USA: Association for Computing Machinery, 2020, pp. 297–302, doi: 10.1145/3339825.3394937.
[26] A. Golinski, R. Pourreza, Y. Yang, G. Sautiere, and T. S. Cohen, "Feedback recurrent autoencoder for video compression," in *Proc. Asian Conf. Comput. Vis.*, 2020, pp. 1–17.
[27] A. Habibian, T. V. Rozendaal, J. Tomczak, and T. Cohen, "Video compression with rate-distortion autoencoders," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7033–7042.
[28] N. Sigger, N. Al-Jawed, and T. Nguyen, "Spatial-temporal autoencoder with attention network for video compression," in *Proc. 17th Int. Joint Conf. Comput. Vis., Imag. Comput. Graph. Theory Appl. Comput. Vis. Theory Appl. (VISAPP)*, vol. 4. Setúbal, Portugal: SCITEPRESS Digital Library, Feb. 2022, pp. 364–371.
[29] S. Sangeeta, P. Gulia, and N. S. Gill, "Flow incorporated neural network based lightweight video compression architecture," *Indonesian J. Electr. Eng. Comput. Sci.*, vol. 26, no. 2, pp. 939–946, May 2022.
[30] S. Wang, Y. Zhao, H. Gao, M. Ye, and S. Li, "End-to-end video compression for surveillance and conference videos," *Multimedia Tools Appl.*, vol. 81, no. 29, pp. 42713–42730, Dec. 2022.
[31] W. Weaver, "The mathematics of communication," *Sci. American*, vol. 181, no. 1, pp. 11–15, 1949.
[32] G. Shi, D. Gao, X. Song, J. Chai, M. Yang, X. Xie, L. Li, and X. Li, "A new communication paradigm: From bit accuracy to semantic fidelity," 2021, *arXiv:2101.12649*.
[33] G. Shi, Y. Xiao, Y. Li, and X. Xie, "From semantic communication to semantic-aware networking: Model, architecture, and open problems," *IEEE Commun. Mag.*, vol. 59, no. 8, pp. 44–50, Aug. 2021.
[34] H. Xie, Z. Qin, G. Y. Li, and B.-H. Juang, "Deep learning enabled semantic communication systems," *IEEE Trans. Signal Process.*, vol. 69, pp. 2663–2675, 2021.
[35] A. Green, "The hard problem of semantic communication," *Philosophia*, vol. 50, no. 3, pp. 1117–1130, Jul. 2022.
[36] A. Li, X. Wei, D. Wu, and L. Zhou, "Cross-modal semantic communications," *IEEE Wireless Commun.*, vol. 29, no. 6, pp. 144–151, Dec. 2022.
[37] H. Xie, Z. Qin, G. Y. Li, and B.-H. Juang, "Deep learning based semantic communications: An initial investigation," in *Proc. GLOBECOM-IEEE Global Commun. Conf.*, Dec. 2020, pp. 1–6.

[38] X. Luo, H.-H. Chen, and Q. Guo, "Semantic communications: Overview, open issues, and future research directions," *IEEE Wireless Commun.*, vol. 29, no. 1, pp. 210–219, Feb. 2022.
[39] Y. Zhang, H. Zhao, J. Wei, J. Zhang, M. F. Flanagan, and J. Xiong, "Context-based semantic communication via dynamic programming," *IEEE Trans. Cognit. Commun. Netw.*, vol. 8, no. 3, pp. 1453–1467, Sep. 2022.
[40] M. Zhang, Y. Li, Z. Zhang, G. Zhu, and C. Zhong, "Wireless image transmission with semantic and security awareness," *IEEE Wireless Commun. Lett.*, vol. 12, no. 8, pp. 1389–1393, May 2023.
[41] A. Li, X. Liu, G. Wang, and P. Zhang, "Domain knowledge driven semantic communication for image transmission over wireless channels," *IEEE Wireless Commun. Lett.*, vol. 12, no. 1, pp. 55–59, Jan. 2023.
[42] E. Erdemir, T.-Y. Tung, P. L. Dragotti, and D. Gündüz, "Generative joint source-channel coding for semantic image transmission," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 8, pp. 2645–2657, Jun. 2023.
[43] M. U. Lokumarambage, V. S. S. Gowrisetty, H. Rezaei, T. Sivalingam, N. Rajatheva, and A. Fernando, "Wireless end-to-end image transmission system using semantic communications," *IEEE Access*, vol. 11, pp. 37149–37163, 2023.
[44] P. Jiang, C.-K. Wen, S. Jin, and G. Y. Li, "Wireless semantic communications for video conferencing," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 1, pp. 230–244, Jan. 2023.
[45] S. Wang, J. Dai, Z. Liang, K. Niu, Z. Si, C. Dong, X. Qin, and P. Zhang, "Wireless deep video semantic transmission," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 1, pp. 214–229, Jan. 2023.
[46] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. 13th Int. Conf. Artif. Intell. Statist.*, vol. 9, Y. W. Teh and M. Titterington, Eds., Sardinia, Italy, 2010, pp. 249–256. [Online]. Available: https://proceedings.mlr.press/v9/glorot10a.html
[47] W. A. C. Fernando, C. N. Canagarajah, and D. R. Bull, "A unified approach to scene change detection in uncompressed and compressed video," *IEEE Trans. Consum. Electron.*, vol. 46, no. 3, pp. 769–779, Jun. 2000.
[48] P. Midtrack. *People Enjoying the Day in a Beach*. Accessed: Feb. 12, 2024. [Online]. Available: https://www.pexels.com/video/people-enjoying-the-day-in-a-beach-3150419/
[49] T. Miroshnichenko. *People Playing Soccer*. Accessed: Feb. 12, 2024. [Online]. Available: https://www.pexels.com/video/people-playing-soccer-6077718/
[50] C-Couple. *A Bowl of Avocados and Vegetables*. Accessed: Feb. 12, 2024. [Online]. Available: https://www.pexels.com/video/a-bowl-of-avocados-and-vegetables-7656166/
[51] Pixabay. *Blue Sky Video*. Accessed: Feb. 12, 2024. [Online]. Available: https://www.pexels.com/video/blue-sky-video-855005/
[52] T. Elliot. *A Couple Walking Towards the Launching Area of the Hot Air Balloons Festival*. Accessed: Feb. 12, 2024. [Online]. Available: https://www.pexels.com/video/a-couple-walking-towards-the-launching-area-of-the-hot-air-balloons-festival-3064025/
[53] D. C. Paduret. *Culturing a Chamomile Flower Plant*. Accessed: Feb. 12, 2024. [Online]. Available: https://www.pexels.com/video/culturing-a-chamomile-flower-plant-3011973/
[54] S. Garenko. *A Girl Running Across a Bridge*. Accessed: Feb. 12, 2024. [Online]. Available: https://www.pexels.com/video/a-girl-running-across-a-bridge-19805236/
[55] V. Singh. *Rangdum Village in Zanskar Valley*. Accessed: Feb. 12, 2024. [Online]. Available: https://www.pexels.com/video/rangdum-village-in-zanskar-valley-19022224/
[56] M. Kilinc. *Nemrut - Bitlis*. Accessed: Feb. 12, 2024. [Online]. Available: https://www.pexels.com/video/nemrut-bitlis-18856748/
[57] M. T. Kirkgoz. *Cold Snow Sea Dawn*. Accessed: Feb. 12, 2024. [Online]. Available: https://www.pexels.com/video/cold-snow-sea-dawn-18051870/
[58] *Old Town Cross. Xiph.org*. Accessed: Apr. 2, 2024. [Online]. Available: https://media.xiph.org/video/derf/
[59] *Controlled Burn. Xiph.org*. Accessed: Apr. 2, 2024. [Online]. Available: https://media.xiph.org/video/derf/
[60] *Ducks Take Off. Xiph.org*. Accessed: Apr. 2, 2024. [Online]. Available: https://media.xiph.org/video/derf/
[61] *Four People. Xiph.org*. Accessed: Apr. 2, 2024. [Online]. Available: https://media.xiph.org/video/derf/
[62] *Galleon. Xiph.org*. Accessed: Apr. 2, 2024. [Online]. Available: https://media.xiph.org/video/derf/
[63] L. Photography. *A Wooden Bridge Built Above Water*. Accessed: Apr. 2, 2024. [Online]. Available: https://www.pexels.com/video/wood-sea-landscape-nature-3971351/
[64] C. Studio. *A Person Writing on the Blueprint*. Accessed: Apr. 2, 2024. [Online]. Available: https://www.pexels.com/video/a-person-writing-on-the-blueprint-7484777/
[65] P. Whelen. *Man Walking on Road Among Pine Trees*. Accessed: Apr. 2, 2024. [Online]. Available: https://www.pexels.com/video/man-walking-on-road-among-pine-trees-5738706/
[66] P. Whelen. *A Person Doing Graffiti Art By Using a Spray Paint*. Accessed: Apr. 2, 2024. [Online]. Available: https://www.pexels.com/video/a-person-doing-graffiti-art-by-using-a-spray-paint-5621707/
[67] A. Podrez. *A Young Man Executing a Kick Flip Skateboard Trick*. Accessed: Apr. 2, 2024. [Online]. Available: https://www.pexels.com/video/a-young-man-executing-a-kick-flip-skateboard-trick-4832083/
[68] N. Barman, N. Khan, and M. G. Martini, "Analysis of spatial and temporal information variation for 10-bit and 8-Bit video sequences," in *Proc. IEEE 24th Int. Workshop Comput. Aided Model. Design Commun. Links Netw. (CAMAD)*, Sep. 2019, pp. 1–6.
[69] A. Wieckowski, J. Brandenburg, T. Hinz, C. Bartnik, V. George, G. Hege, C. Helmrich, A. Henkel, C. Lehmann, C. Stoffers, I. Zupancic, B. Bross, and D. Marpe, "VVenC: An open and optimized VVC encoder implementation," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Jul. 2021, pp. 1–2.
[70] S. Tomar, "Converting video formats with FFmpeg," *Linux J.*, vol. 2006, no. 146, p. 10, 2006.
[71] Q. Huynh-Thu and M. Ghanbari, "The accuracy of PSNR in predicting video quality for different video scenes and frame rates," *Telecommun. Syst.*, vol. 49, no. 1, pp. 35–48, Jan. 2012.
[72] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[73] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Mahohara. (Apr. 2017). *Toward a Practical Perceptual Video Quality Metric*. Accessed: Feb. 12, 2024. [Online]. Available: https://netflixtechblog.com/toward-a-practical-perceptual-video-quality-metric-653f208b9652
[74] G. Bjontegaard, *Calculation of Average Psnr Differences Between Rd-curves*, document VCG-M33, 2001. [Online]. Available: https://www.itu.int/wftp3/av-arch/video-site/0104_Aus/VCEG-M33.doc
[75] S. Acharjee, S. Chakraborty, W. B. A. Karaa, A. T. Azar, and N. Dey, "Performance evaluation of different cost functions in motion vector estimation," *Int. J. Service Sci., Manag., Eng., Technol.*, vol. 5, no. 1, pp. 45–65, Jan. 2014.
[76] S. Acharjee, D. Biswas, N. Dey, P. Maji, and S. S. Chaudhuri, "An efficient motion estimation algorithm using division mechanism of low and high motion zone," in *Proc. Int. Mutli-Conference Autom., Comput., Commun., Control Compressed Sens. (iMacs)*, Mar. 2013, pp. 169–172.
[77] W. Ahmad, H. Mahdavi, and I. Hamzaoglu, "An efficient versatile video coding motion estimation hardware," *J. Real-Time Image Process.*, vol. 21, no. 2, p. 25, Apr. 2024.
[78] M. Yang, C. Bian, and H.-S. Kim, "Deep joint source channel coding for WirelessImage transmission with OFDM," 2021, *arXiv:2101.03909*.

PRABHATH SAMARATHUNGA (Graduate Student Member, IEEE) received the B.Sc. degree (Hons.) in software engineering from the University of Plymouth, U.K., in 2021. He is currently pursuing the Ph.D. degree in semantic communication-based video streaming for M2M communication with the Department of Computer and Information Sciences, University of Strathclyde, U.K. His research interests include semantic communication, video coding, video streaming in error-prone channels, and machine learning.

YASITH GANEARACHCHI (Graduate Student Member, IEEE) received the B.Sc. degree (Hons.) in electrical and electronic engineering from the University of Peradeniya, Sri Lanka, in 2008, the M.B.A. (Merit) degree from the University of Sri Jayewardenepura, Sri Lanka, in 2019, and the P.M.B.A.D.T. (Merit) degree from Asian Institute of Technology, Thailand, in 2022. He is currently pursuing the Ph.D. degree in computer and information sciences with the University of Strathclyde, U.K. Following 15 years of telecommunications industry experience during the Ph.D. degree with Sri Lanka Telecom PLC.

THANUJ FERNANDO (Student Member, IEEE) is currently pursuing the B.Sc. degree in computer science with the Department of Computer and Information Sciences, University of Strathclyde, U.K. His research interests include machine learning, vision processing, and communications.

ADHURAN JAYASINGAM (Member, IEEE) received the B.S. (Eng.) degree (Hons.) in electronics engineering and the M.Eng. degree in microelectronics and embedded systems from Asian Institute of Technology, Thailand, in 2015 and 2017, respectively, and the Ph.D. degree from the Center for Vision, Speech, and Signal Processing, University of Surrey, in 2022. He worked as a Lecturer (Probationary) at the University of Jaffna, Sri Lanka, from July 2018 to September 2018. He also worked for BBC Research and Development as a R&D Engineer. He is currently a Visiting Researcher with Kingston University, U.K. He has authored several technical papers on video coding, video quality assessment, and video streaming technologies. His current research focus is in the field of video coding and video quality assessments.

INDIKA ALAHAPPERUMA (Graduate Student Member, IEEE) received the B.Sc. degree (Hons.) in electronics and telecommunication engineering from the University of Moratuwa, Sri Lanka, in 2003, and the M.B.A. degree from the University of Sri Jayewardenepura, Sri Lanka, in 2016. He is currently pursuing the Ph.D. degree in computer and information sciences with the University of Strathclyde, U.K. Following 20 years of telecommunications industry experience during the Ph.D. degree.

ANIL FERNANDO (Senior Member, IEEE) received the B.Sc. degree (Hons.) in electronics and telecommunication engineering from the University of Moratuwa, Sri Lanka, in 1995, the M.Sc. degree (Hons.) in communications from Asian Institute of Technology, Bangkok, Thailand, in 1997, and the Ph.D. degree in computer science (video coding and communications) from the University of Bristol, U.K., in 2001. He is currently a Professor in video coding and communications with the Department of Computer and Information Sciences, University of Strathclyde, U.K., where he leads the Video Coding and Communication Research Team and also a Visiting Professor with the Center for Vision, Speech and Signal Processing (CVSSP), University of Surrey, U.K. He has been working with all major EU broadcasters, BBC, and major European media companies/SMEs in the last decade to provide innovative media technologies for British and EU citizens. He has graduated more than 110 Ph.D. students and is currently supervising 20 Ph.D. students. He has worked on major national and international multidisciplinary research projects and led most of them. He has published more than 430 papers in international journals and conference proceedings and has published a book on 3D video broadcasting. His main research interests include video coding and communications, machine learning, artificial intelligence, semantic communications, signal processing, networking and communications, interactive systems, resource optimization in 6G, distributed technologies, media broadcasting, and quality of experience.