



Optimizing Video Transcoding Using Generative Adversarial Networks and Autoencoders: A Semantic-Aware Compression Framework with Quality Prediction on Authentically Distorted Video

Venkateswara Reddy B^{*}, Bala K

Department of Computer Science and Engineering, School of Computing, Bharath Institute of Science and Technology (BIST), Chennai 600073, India

Corresponding Author Email: venkateswarareddyb17@gmail.com

Copyright: ©2026 The authors. This article is published by IETA and is licensed under the CC BY 4.0 license (<http://creativecommons.org/licenses/by/4.0/>).

<https://doi.org/10.18280/jesa.590919>

ABSTRACT

Received: 5 July 2026

Revised: 10 September 2026

Accepted: 23 September 2026

Available online: 30 September 2026

Keywords:

video transcoding, Generative Adversarial Networks, autoencoders, video quality assessment, semantic compression, edge computing

The growing number of high-quality users generated videos requires compression-efficient transcoding architectures that can keep the human-perceived visual quality. Traditional block-based codes have intrinsic problems in coding in-the-wild videos with authentic distortions, often causing prominent degradation artifacts at low bitrates. In this paper, we aim to overcome these drawbacks by introducing an end-to-end framework Semantic-Aware GAN-Autoencoder Transcoding (SAGAT) that combines learned compression with adversarial perceptual restoration. SAGAT has the advantages that it is trained and evaluated directly on LIVE Video Quality Challenge (VQC) Database, which has complex and commingled authentic distortions, unlike existing methods trained on clean synthetic data. To learn quality-aware representations explicitly, we propose architecture with a Quality Prediction Head (QPH) in the latent space supervised by the crowd-sourced Mean Opinion Scores (MOS). The extensive evaluations show that SAGAT can compress 38.2% more than H.265/HEVC standard while still having a strong perceptual quality with high VMAF score of 85.2 and low LPIPS score of 0.081. Moreover, the embedded quality predictor can closely correlate with the human subjective assessments (SROCC: 0.7645) and therefore shows good performance over classical No-Reference Video Quality Assessment models. Semantic-guided bit allocation and temporal coherence constraints are adopted to offer a feasible and efficient solution to be integrated into the current edge-cloud video delivery networks.

1. INTRODUCTION

With the rapid growth of video streaming [1] services, video traffic has become the predominant type of internet traffic and is now the primary consumer of global internet network bandwidth [2]. Platforms that enable user-generated content are always in need of efficient ways to deliver a high-quality video over a diversity of devices, in very unpredictable network conditions. The basic operation that makes this adaptive delivery paradigm of videos [3] possible is called video transcoding, that is, the conversion of compressed video from one format, resolution, and/or bitrate to another [4]. The traditional transcoding pipelines, however, are based on standardized codes like H.264/AVC and H.265/HEVC that have inherent limitations creating strain to modern streaming infrastructure [5] and quality of experience (QoE) [6] for the end-user [7]. The main methods of compression that conventional video codecs [8] use are deterministic, block-based motion compensation [9], transform coding (e.g., discrete cosine transform) [10] and rate-distortion optimization [11]. Though these methods have been the basis of digital video distribution, they have major drawbacks in the use in highly constrained modern environments. First, their

home-grown encoding methods (block partitioning, motion vector estimation, quantization heuristics) do not have any semantic knowledge of the visual content, but only rigid spatial structures [12]. The exhaustive mode evaluation needed to obtain the optimum encoding parameters is too costly for real time use on edge devices [13]. Secondly, these codecs have at the high compression ratios needed to cope with the varying bandwidths of the cells inevitably introduced visually objectionable artefacts such as blocking, ringing and colour bleeding [14]. These artefacts greatly reduce the viewing experience especially on high-quality screens where spatial fidelity is a key factor. Third, it is the traditional decode-then-reencode transcoding pipeline that is intrinsically lossy, mathematically destructive and discards valuable residual information from the source bit stream, leading to monotonic quality degradation at every transcoding in the distribution system [15]. A basic constraint that aggravates these transcoding problems is the methodology used for assessing the quality of videos in the existing research on video quality assessment (VQA) [16]. Most of the No-Reference (NR) VQA models and deep-learning [17] transcoders are exclusively based on distortions synthetically generated in a database [18]. These traditional databases have

a very restricted diversity of content, laboratory resolution and are distorted with a very limited number of mathematically simulated distortions (e.g., uniform Gaussian noise, artificial blur or fixed-QP compression) [19]. As a result, these datasets are not sufficiently representative of real-world (in-the-wild) videos. Modern user-generated content is taken as images under a wide range of imaging conditions, with heterodox imaging optics and with authentic and complicated distortions, which are impossible to simulate accurately, such as sensor noise, poor exposure and hand-shake blur (or jitter). Thus, transcoding systems optimized on synthetic content show a significant degradation on real content [20]. Such transcoding problems are compounded by the evolution of the architecture of modern content distribution networks (CDNs) and edge computing models [21]. The idea of edge computing is to move transcoding and caching to the edge of the network to significantly decrease the latency in transmission and congestion in the backhaul network [22]. Recent works have shown that the edge of the network can be optimally exploited, through the combination of caching and transcoding, using deep reinforcement learning and federated caching architectures [23]. These optimizations, however, are mostly for the delivery pipeline routing and caching logic and not for the transcoding operation itself, and do not affect the mathematical quality-compression trade-off. Even with an intelligent delivery network, an algorithm for perceptual fidelity is still too rigid to dynamically adjust its fidelity. Recent developments in deep neural networks give transformative ways to overcome these classical limitations [24-25]. Generative Adversarial Networks (GANs) trained using the minimax adversarial learning framework have been shown to be very powerful in generating very photo-realistic textures, as well as recovering very degraded image data [7]. Adversarial training gives a mathematically sound way of steering the generated output towards the theoretical limit of perceptual quality, which is much better than the blurry output obtained with traditional Mean Squared Error (MSE) loss optimization [9]. At the same time, Convolutional AutoEncoders have demonstrated significant potential for learned compression which aims to learn very compact non-linear latent representations to efficiently encode the whole semantic structure of visual data without any strict block partition [6, 11]. GANs for perceptual quality enhancement and autoencoders for learned semantic compression seem to be a great opportunity for rethinking the video transcoding from scratch.

But the core issue is not just to fuse the compression that has been learned with the generative restoration. Current GAN-based compression/restoration approaches are mostly adversarial based to enhance perceptual quality of the reconstructed content, while learned compression methods are mostly developed to gain compact representations and reduce reconstruction distortion. Semantic-Aware GAN-Autoencoder Transcoding (SAGAT), on the other hand, considers perceptual video transcoding as a problem of content adaptive representation where the importance of the semantic content directly affects the compression process. Additionally, the learned latent representation is constrained with human subjective Mean Opinion Scores (MOS), which put an explicit relationship between compression representation and perceived visual quality. Also, since the goal is to create a video and not a single image, temporal consistency is also taken into account, to avoid perceptual improvements that may produce inconsistencies and flickering between frames. It thus

departs from the usual proposal of a combination of an autoencoder and a GAN in various ways: it deals with what information to keep, how to represent it, how to use human perception to guide the representation, and how perceptual quality should not change over time.

Instead of using handcrafted heuristics, we suggest to learn one-to-one transcoding function end-to-end. We hypothesize that the same neural network can learn semantically meaningful representations with autoencoding, recover perceptual high-frequency details by learning adversarial terms, and most importantly, learn to predict perceptual quality scores which are strongly correlated with human judgements on “authentically” distorted videos. This paper makes its major contribution in the following systematic manner:

- a. We introduce a Semantic-Aware GAN-Autoencoder framework to tackle the problem of perceptual video transcoding in an end-to-end way.
- b. Semantic guided bit allocation strategy allocates bits to the regions that are important to the human visual system.
- c. Feature preservation is maintained with an embedded quality prediction module which is oriented towards human perception.
- d. The framework is extensively tested on LIVE VQC dataset and provides better performance under real world distortions.

2. LITERATURE SURVEY

Traditional video transcoding approaches have been based on cascaded configuration or systems that make use of the input bitstream information to speed up the re-encoding process as reported by Choi et al. [4]. The quality-complexity trade-off in such systems has been well solved by machine learning; for example, Escribano et al. [12] suggested to replace exhaustive mode evaluations with decision tree classifiers, for a dramatic reduction of transcoding time. These classical ML methods demonstrated that it is possible to replace handcrafted heuristic methods with learned decision functions but are intrinsically constrained by the block-partitioning structure of the standards that they are used with, such as H.264 and H.265. This means rather than a complete rethinking of the mathematical method of compression, they make the existing algorithms faster. At the same time, the needs of modern distribution networks have led to the implementation of edge computing for optimizing the delivery of video. Deep Reinforcement Learning (DRL) is recently widely used to handle caching, adaptive bitrate streaming and transcoding jointly at the edge of the network, as done by Yang and Xiao [1] and Luo et al. [8]. Furthermore, there are techniques that combine super-resolution with edge caching, like those proposed by Cao and Gao [13] that minimizes the cost of content delivery to the edge when using super-resolution, and multi-agent and federated learning approaches by Shi et al. [2] and Sinthia et al. [5] have validated the significance of content-aware edge caching in enhancing overall network efficiency. Most of these network-level systems, however, are mainly oriented towards optimizing the delivery route and managing the caches and basically assume that the underlying video compression architecture is an unoptimized black box. Recently, compression has seen a paradigm change in favour of deep learning architectures, aiming to radically enhance compression. Autoencoders will

naturally learn a compact latent representation, and more sophisticated approaches by Hyder and Asif [6] use low rank temporal constraints and joint optimization of latent codes and generator weights, to result in high compression rates. Incorporation of attention mechanism inside these autoencoders, as shown by Song et al. [11] further enables the models to learn to focus on informative regions of spatiotemporal data to reduce the entropy of the transmitted bitstream. While all these methods have demonstrated the capabilities of learned spatial compression, they often result in blurry reconstruction images because they are based on deterministic MSE loss functions. In parallel to the development of autoencoders, GANs have also shown outstanding ability in visual synthesis and perceptual quality optimization which goes beyond the traditional MSE based methods, as shown by Liu et al. [7]. The combination of spatial CNN with temporal LSTM and the use of multi-component loss functions, as proposed by Darwich and Bayoumi [9] has contributed greatly in enhancing the visual satisfaction. In addition, semantic video coding based on GANs, which assigns more bitrate to the foreground faces in the scene and aggressively compress the background, is an appealing method to allocate bitrate, as demonstrated by Galteri et al. [14]. However, it is important to have powerful No-Reference Video Quality Assessment (NR-VQA) models to evaluate the success of these generative enhancements. While creating the LIVE Video Quality Challenge (VQC) Database, huge deficiencies in current hand-crafted VQA metrics (such as VIIDEO, NIQE and BRISQUE) were encountered against complex, commingled authentic distortions, as pointed out by Mittal et al. [17] and Sinno and Bovik [18, 19].

To capture the variations of video content and enhance the compression efficiency, Tang et al. [21] introduced HANeRV, a hierarchical implicit neural representation (INR) framework that optimizes network structures adaptively at both the sequence and frame levels, as well as at the spatial level. The framework features dynamic architecture-level adjustment, dynamic frame-level adjustment, and hierarchical structural adaptation, which achieves state-of-the-art performance among INR-based video compression methods and surpasses the H.266/VVC anchor on various datasets. Likewise, Chen et al. [22] proposed a spatio-temporal adaptive deep video compression scheme to adaptively compress the video content by varying spatial resolution and temporal frame rate. The method improves the rate-distortion performance, compared to the others, by eliminating both spatial and temporal redundancies using a learned adaptation network, which results in bitrate savings in terms of PSNR and MS-SSIM. The studies reveal that in order to build more efficient learned video compression systems, adaptive neural representations and content-aware spatio-temporal optimization of the representation seem to have significant potential.

All these developments, however, still leave an important gap in perceptual video transcoding. Currently learned compression, generative restoration, semantic video coding and video delivery techniques that mostly solve various aspects of the problem. First, learned compression methods aim to optimize compact representations and reconstruction quality of the image, but they do not explicitly aim at optimizing the subjective quality of the image as perceived by the human observer. Second, the GAN-based restoration methods have the ability to enhance perceptual realism but are originally developed for image/framed based restoration and not for deciding which part should have more representation

capacity when compressing. Third, typically semantic allocation is not associated with MOS-supervised latent learning and temporally consistent generative restoration. Lastly, most deep-learning methods are tested on synthetic or synthetically corrupted videos, while the videos created by users have complex real-world distortions. The gaps inspire SAGAT, where semantic-aware compression, human-perception-guided representation learning, perceptual restoration and temporal consistency are all seen as related tasks that can be brought together in a single framework of the transcoding of a video.

3. PROPOSED WORK

The proposed SAGAT framework is an end-to-end deep learning framework to provide efficient video compression with perceptual quality. The framework is composed of a semantic-aware autoencoder, a quantization parameter prediction head, a relativistic generative adversarial network (GAN), and temporal consistency optimization, and is implemented on a single pipeline. Firstly, the encoder extracts compact latent representations with the help of masking of semantic important images, while the decoder recovers compressed frames. A generator is then used to further process the reconstructed images and add back high-frequency information, and a relativistic discriminator enhances the perceptual realism. Finally, temporal coherence is enforced across consecutive frames, to reduce flickering artefacts, by optimizing using optical flow. The whole architecture is presented in Figure 1.

The SAGAT system of training has three successive stages. The encoder, decoder and quantization parameter prediction head are optimized in phase 1 by minimizing reconstruction, perceptual, rate and MOS prediction losses to learn compact semantic representations. During Phase 2, a Relativistic GAN is used to improve reconstructed frames in an adversarial fashion along with autoencoder optimization. In Phase 3, temporal coherence loss is added to ensure the consistency of a video between consecutive frames, and thus achieve the visually realistic, temporally stable and compression-efficient video transcoding.

3.1 System architecture overview

To increase the compression efficiency without compromising perceptual video quality, an end-to-end video transcoding architecture, called SAGAT, is proposed. The framework is made up of five modules that are integrated. The Semantic Analyzer aims to detect important regions visually to make compression based on content. The Autoencoder Compressor (AEC) encodes and reconstructs the video frames and allocates more resource Quality Prediction Head (QPH) to the perceptually significant areas. The QPH helps the model retain features that are important to human visual perception. The GAN Restorer module improves the quality of the reconstructed frames by incorporating details of fine textures and structures, and the Temporal Consistency Module ensures smoother transitions between sequential frames and mitigates flickering artifacts. These modules in combination give an efficient compression with high visual quality and stable temporal consistency for real-world video transcoding applications. In the end-to-end data flow, the following mathematical function describes data flow during inference.

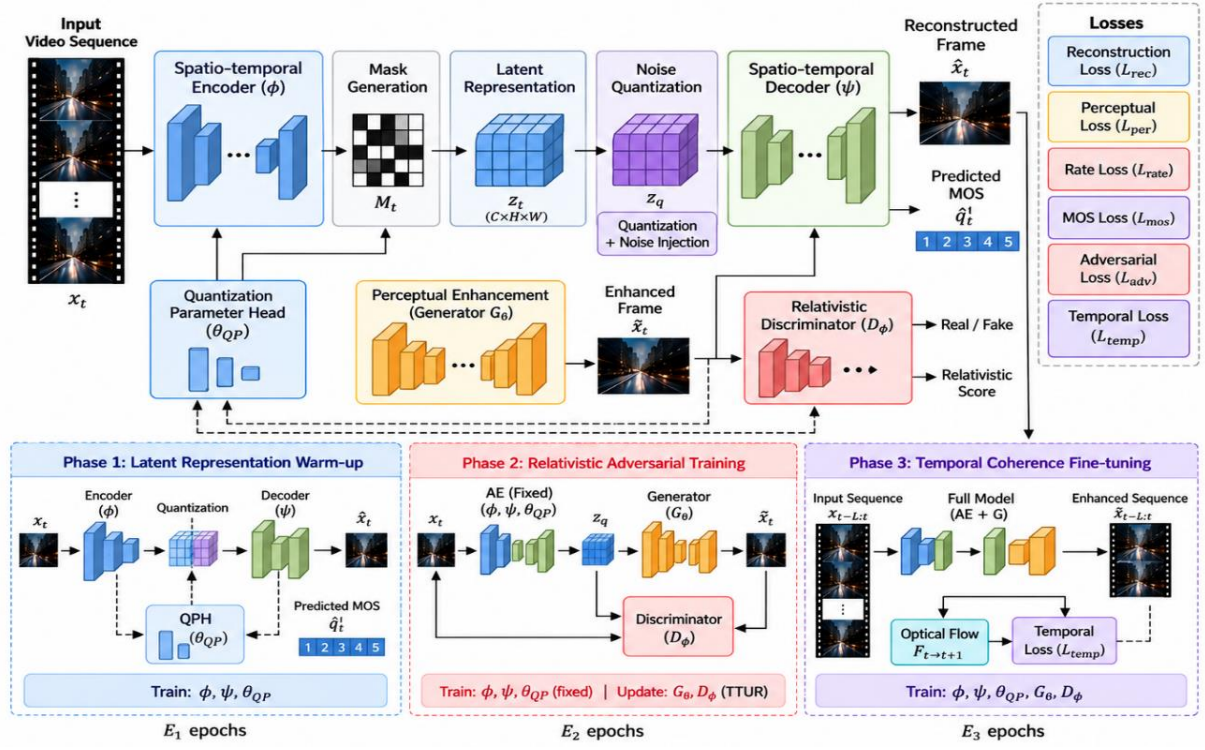


Figure 1. Proposed Semantic-Aware GAN-Autoencoder Transcoding (SAGAT) architecture

$$x_t^{SA} \xrightarrow{E_\phi} M_t \xrightarrow{\text{split}} \begin{cases} f_{QP} \wedge \rightarrow q_t \text{ (predicted MOS)} \\ D_\psi \wedge G_\theta \sim \rightarrow x_t \rightarrow \tilde{x}_t \text{ (transcoded output)} \end{cases} \quad (1)$$

SAGAT is implemented as an automated content-adaptive transcoding pipeline at system level, where the system decides on the processing strategy based on the content of the incoming video. The Semantic Analyzer automatically identifies perceptually important regions for each input frame then creates a semantic mask. This mask is then given to the Autoencoder Compressor to adjust the representation of spatial regions in relation to their perceptual importance. As a result of this, the compression process doesn't compress all the spatial regions equally and semantically significant regions are compressed with a higher representation priority, whereas less significant background regions can be compressed more compactly. This is achieved by the encoder formulation described in Section 3.2. The automated processing is then carried out in the QPH, GAN Restorer, and Temporal Consistency Module. The QPH gives a quality-aware learning signal, being able to predict the perceptual quality from the learned representation, and constraining the latent space with the MOS information. The GAN Restorer then refines the reconstructed frames, and the Temporal Consistency Module exploits the optical-flow data between the consecutive frames to remove temporal inconsistency and flickering. Therefore, content analysis and representation, content compression and quality, perceptual restoration and content refinement, are automatically executed in turn, to generate the output.

3.2 Semantic Analyzer formulation

The Semantic Analyzer is a standalone preprocessing network which maps the input spatial domain to a probability-bounded saliency domain. The network produces the binary-thresholder saliency maps $M_t \in \{0,1\}^{H \times W}$, in which a pixel $M_t(i,j) = 1$ indicates high perceptual priority of the pixel. We

use a simple U-Net-like segmentation network, trained with initial weights on gigantic saliency detection datasets, which is only moderately compute-intensive. Spatial mapping is defined as:

$$M_t = \sigma(f_{SA}(x_t; \theta_{SA})) \quad (2)$$

where, $\sigma(\cdot)$ is a sigmoid activation function that is severely thresholded at a probability level of 0.5. For the LIVE VQC dataset, which contains a lot of commingled distortions, the network has three separate functional pathways to calculate the joint probability of saliency: One of them is a face-detection cascade essential for user generated mobile video content. A TD pathway to isolate foreground kinematics. An 'spatial high-frequency filter bank' to maintain critical texture boundaries under large amounts of compression. A binary mask M_t then creates a hard, non-uniform spatial binary spatial matrix of the frame:

Foreground ($\mathcal{R}_{FG}$): $\mathcal{R}_{FG} = \{(i,j): M_t(i,j) = 1\}$. This domain is given priority in channel allocation during encoding such that high perceptual fidelity is obtained.

Background ($\mathcal{R}_{BG}$): $\mathcal{R}_{BG} = \{(i,j): M_t(i,j) = 0\}$. The spatial decimation/quantization of this domain is done aggressively so as to maximize the overall compression ratio.

3.3 Autoencoder compressor

The core compression engine is based on an asymmetric encoder-decoder based architecture (E_ϕ and D_ψ). Encoder Formulation: Encoder will project the input space frames to achieve a lower dimensional latent representation. To use the semantic guidance, the mask M_t is used as a spatial attention modulator before the deep convolutional cascade:

$$z_t = E_\phi(x_t, M_t) = f_{enc}(x_t \odot (1 + \lambda M_t); \phi) \quad (3)$$

In this case $\odot$ is the Hadamard (elementwise) product, and

λ is a tenable scalar parameter to control the relative magnitude of the gradients of foreground regions compared to the background. The network consists of five stride convolutional blocks with PReLUs and residual bypass connections, including an aggressive down sampling of the spatial dimension to obtain a small latent tensor $z_t \in \mathbb{R}^{C_z \times \frac{H}{16} \times \frac{W}{16}}$.

Standard backpropagation is not possible over a hard rounding function - Latent Quantization. To make it possible to use gradient flow end-to-end, we make the approximation of quantization in the training phase by adding additive uniform noise:

$$\hat{z}_t = z_t + \epsilon, \epsilon \sim \mathcal{U}(-0.5, 0.5) \quad (\text{Training Phase}) \quad (4)$$

This is changed to deterministic integer rounding during inference:

$$\hat{z}_t = \lfloor z_t \rfloor \quad (\text{Inference Phase}) \quad (5)$$

Decoder Formulation: The decoder D_ψ is like encoder, but with transposed convolutions and pixel-shuffle layers to rebuild the original baseline spatial matrix:

$$\hat{x}_t = D_\psi(\hat{z}_t; \psi) \quad (6)$$

3.4 Quality Prediction Head

In traditional autoencoders, no subjective perception of a human is considered, and the fidelity of the pixels are optimized according to L1/L2 loss only. To fill this gap we include the QPH $f_{QP}(\odot; \theta_{QP})$. This sub-network is designed to map the very non-linear latent bottleneck vector to an MOS prediction in the form of a single scalar value:

$$\hat{q}_t = f_{QP}(\hat{z}_t; \theta_{QP}) \quad (7)$$

From an architectural point of view, the QPH uses a global spatial average pooling to reduce spatial dimensions, and then a dense stack of three fully connected layers, each with LeakyReLU activations and heavy dropout ($p = 0.3$) to avoid overfitting too much on such small VQA datasets. Joint backpropagation from the QPH into the encoder weights ϕ mathematically makes the latent space z_t carry explicit knowledge of the true distortion parameters (such as the intensity of motion blur, the variance of the sensor noise, etc.) that affect human visual perception.

3.5 GAN Restorer dynamics

The autoencoder output $\hat{x}_t$ is inherently blurry because of the blurriness inherent in deterministic L1 reconstruction loss at high compression rates. GAN Restorer can fix this with the help of Generator G_θ and Discriminator D_ω .

Generator G_θ : The generator is made up of 16 continuous Residual-in-Residual Dense Blocks (RRDB), which can be regarded as a deep spatial refinement filter. All Batch Normalization layers are explicitly removed to avoid artifacts due to different resolution normalizations in batches. The formulation is an addition of spaces, a reminder of space, a residue of space:

$$\tilde{x}_t = \hat{x}_t + G_\theta(\hat{x}_t; \theta) \quad (8)$$

Discriminator D_ω : We make use of a Relativistic Average Discriminator. The relativistic discriminator does not predict if an image is real or fake, but rather the probability that the real image x_t is more realistic than the generated image $\tilde{x}_t$, relative to each other. This relativistic formulation offers significantly more robust gradients in the very dynamic optimization space of real world "in-the-wild" distortions.

3.6 Temporal consistency module

The temporal divergence is always present in frame independent adversarial processing and is very noticeable to the human eye as flickering. We do this by computing deep optical flow $F_{t \rightarrow t-1}$ between successive enhanced frames using a powerful and pre-trained RAFT (Recurrent All-Pairs Field Transforms) network. It is the temporal loss that is extremely punitive on spatial deviation along the given flow vectors:

$$L_{temporal} = \frac{1}{|\Omega|} \sum_{(i,j) \in \Omega} \|\tilde{x}_t(i,j) - \text{warp}(\tilde{x}_{t-1}, F_{t \rightarrow t-1})(i,j)\|_1 \quad (9)$$

where, $\text{warp}(\odot)$ represents a warping operation in the spatial domain which is differentiable.

3.7 Composite optimization objective

The framework is optimized on a very complex, multi-objective landscape with six different mathematical loss components:

$$L_{total} = \lambda_1 L_{rec} + \lambda_2 L_{perc} + \lambda_3 L_{adv} + \lambda_4 L_{temporal} + \lambda_5 L_{rate} + \lambda_6 L_{mos} \quad (10)$$

Reconstruction Loss (L_{rec}): A spatially weighted L1 norm that enforces the basic geometric and colour constraint with the semantic masks that will heavily punish the foreground errors:

$$L_{rec} = \frac{1}{N} \sum_t [w_{FG} \|\tilde{x}_t - x_t\|_1^{FG} + w_{BG} \|\tilde{x}_t - x_t\|_1^{BG}] \quad (11)$$

Perceptual Loss (L_{perc}): A deep feature matching loss, based on deep intermediate layers of the pre-trained VGG-19 network, which is crucial for texture consistency.

Dversarial Loss (L_{adv}): The component of the generator loss in the spatial output distribution, which is being pushed towards the target natural image manifold, is measured in a relativistic manner.

Temporal Loss ($L_{temporal}$): Optical flow warping constraint from Section 3.6.

Rate Loss (L_{rate}): An entropy estimation constraint on the latent codes to explicitly push for data compression.

MOS Quality (L_{mos}): A strong version of the Smooth L1 (Huber) regression loss which minimises the gap between the QPH prediction and the ground-truth crowdsourced MOS scores from the dataset:

$$L_{mos} = \text{SmoothL1}(\hat{q}_t, q_{mos}) \quad (12)$$

Empirical grid search established the optimal loss weighting hyperparameters as:

$$\lambda_1 = 1.0, \lambda_2 = 0.1, \lambda_3 = 0.005, \lambda_4 = 0.5, \lambda_5 = 0.01, \lambda_6 = 0.1.$$

3.8 Optimization algorithm

The end-to-end training of our SAGAT architecture is a non-convex minimax optimization problem, which tries to minimize the objective function from the point of view of the encoder-decoder, quality head and generator, while at the same time maximizing it from the perspective of the discriminator:

$$\min_{\{\phi, \psi, \theta_{QP}, \theta\}} \max_{\omega} \mathbb{E}_{x \sim \mathcal{X}, q \sim \mathcal{Q}} [L_{total}(x, q; \phi, \psi, \theta_{QP}, \theta, \omega)] \quad (13)$$

We use a progressive three-phase learning strategy outlined in Algorithm 1 to ensure a stable convergence to unreferenced videos that are authentic.

Algorithm 1: SAGAT Progressive Training	
Inputs:	Authentically distorted video dataset $\mathcal{X}$, corresponding MOS labels $\mathcal{Q}$, Learning Rates α, β , Phase Epochs E_1, E_2, E_3 .
Output:	Optimized inference network parameters $\{\phi, \psi, \theta_{QP}, \theta\}$.
1.	Initialize all parameters $\{\phi, \psi, \theta_{QP}, \theta, \omega\}$ with Xavier Uniform initialization.
2.	Phase 1: Latent Representation Warm-up
3.	for epoch $\leftarrow 1$ to E_1 do
4.	for batch $\{x_t, q_{MOS}\}$ in $\mathcal{X}, \mathcal{Q}$ do
5.	Compute masks M_t , latents z_t , and noise-quantized latents $\hat{z}_t$ (2,3,4)
6.	Compute baseline reconstruction $\hat{x}_t$ and predicted MOS $\hat{q}_t$ (6,7)
7.	Compute Phase 1 loss: $L_1 \leftarrow \lambda_1 L_{rec} + \lambda_2 L_{perc} + \lambda_5 L_{rate} + \lambda_6 L_{mos}$ (11,12)
8.	Update AE and QPH weights: $\{\phi, \psi, \theta_{QP}\} \leftarrow \{\phi, \psi, \theta_{QP}\} - \alpha \nabla L_1$.
9.	end for
10.	end for
11.	Phase 2: Relativistic Adversarial Training

12.	for epoch $\leftarrow 1$ to E_2 do
13.	for batch $\{x_t, q_{MOS}\}$ in $\mathcal{X}, \mathcal{Q}$ do
14.	Execute AE forward pass to obtain baseline $\hat{x}_t$ (2-6)
15.	Generate perceptually enhanced frame: $\tilde{x}_t \leftarrow \hat{x}_t + G_{\theta}(\hat{x}_t)$ (8)
16.	Compute Relativistic Discriminator loss $L_D(x_t, \tilde{x}_t)$;
17.	update $\omega \leftarrow \omega + \beta \nabla L_D$ using TTUR
18.	Compute generator loss: $L_G \leftarrow L_1 + \lambda_3 L_{adv}$
19.	Update parameters: $\{\phi, \psi, \theta_{QP}, \theta\} \leftarrow \{\phi, \psi, \theta_{QP}, \theta\} - \alpha \nabla L_G$ (13)
20.	end for
21.	end for
22.	Phase 3: Temporal Coherence Fine-tuning
23.	for epoch $\leftarrow 1$ to E_3 do
24.	for sequence batch $\{x_{1..T}\}$ in $\mathcal{X}$ do
25.	Execute full forward pass to obtain enhanced sequence $\{\tilde{x}_{1..T}\}$ (1)
26.	Extract optical flows $F_{t \rightarrow t-1}$ and compute Temporal Loss $L_{temporal}$ (9)
27.	Compute composite optimization objective: $L_{total} \leftarrow L_G + \lambda_4 L_{temporal}$ (10)
28.	Perform joint update on all weights: $\Theta \leftarrow \Theta - \alpha \nabla L_{total}$
29.	end for
30.	end for

4. RESULTS AND DISCUSSION

The implementation-level details of the major modules, such as the Semantic Analyzer, Autoencoder Compressor, QPH, GAN Restorer and Temporal Consistency Module are summarized in Table 1. The key points of each table are the primary layer/block structure, the size and stride of the convolutional kernel, the channel dimensions, the latent representation, and any crucial module-specific parameters. The configuration of training (optimization strategy and hyperparameters of the most important ones) is also summarized to enable independent implementation and comparison.

Table 1. Architectural and implementation specifications of Semantic-Aware GAN-Autoencoder Transcoding (SAGAT)

Module	Architecture	Kernel / Stride	Channels / Dimensions	Parameters / Key Settings
Semantic Analyzer	U-Net-like saliency network + face + TD pathways	3×3 , stride 2	$3 \rightarrow 64 \rightarrow 128 \rightarrow 256 \rightarrow 512$	Pretrained initialization; Sigmoid; threshold = 0.5
Encoder	5 Conv blocks + PReLU + residual bypass	$5 \times 5 / 2$	$3 \rightarrow 64 \rightarrow 128 \rightarrow 192 \rightarrow 256 \rightarrow 320$; $256^2 \rightarrow 8^2$	Latent $320 \times 8 \times 8320 \times 8 \times 8$
Quantization	Additive uniform-noise approximation	—	$320 \times 8 \times 8320 \times 8 \times 8$	U(-0.5,0.5)U(-0.5,0.5) training; rounding inference
Decoder	Transposed Conv + Pixel Shuffle	$5 \times 5 / 2$	$320 \rightarrow 256 \rightarrow 192 \rightarrow 128 \rightarrow 64 \rightarrow 3$; $8^2 \rightarrow 256^2$	PReLU; RGB reconstruction
QPH	GAP + 3 FC layers	—	$320 \rightarrow 128 \rightarrow 64 \rightarrow 1$	LeakyReLU; dropout = 0.30; MOS output
GAN Restorer	RRDB generator + relativistic discriminator	3×3	64 feature channels	16 RRDBs; BN removed; spectral norm + GP-0
Temporal	Pretrained RAFT + bilinear warping	—	Flow: 2 channels	Consecutive-frame consistency; frozen RAFT
Training	3-stage progressive optimization	—	Batch = 8	Adam; LR 10^{-4} to 10^{-5}

Note: QPH = Quality Prediction Head; GAN = Generative Adversarial Network; RRDB = Residual-in-Residual Dense Blocks.

4.1 Dataset characteristics

The LIVE VQC dataset consisted of 70% training videos (410), 10% validation videos (58) and 20% test videos (117). The splitting was done to ensure uniform content, device, distortion and subjective quality score distributions among three subsets. All the frames of a single video sequence have been placed in one subset only to avoid frame-level data leakage. The split was not "device disjoint" so videos recorded with the same device could be in different subsets. The robustness and stability of the proposed framework was tested by running the experiments with several random seeds and reporting the performance as mean $\pm$ STD (standard deviation) for the repeated runs.

4.2 Implementation specifications

The SAGAT architecture is instantiated in PyTorch 2.0. Optimization is executed on an NVIDIA A100 GPU (40 GB VRAM) utilizing the Adam optimizer ($\beta_1 = 0.9, \beta_2 = 0.999$). To enforce discriminator stability, Spectral Normalization is aggressively applied to all convolutional layers within D_ω , accompanied by a zero-centered Gradient Penalty (GP-0, $\gamma = 10$). Training patches of 720×480 are dynamically cropped to manage resolution heterogeneity.

PyTorch 2.0 comes with the Nvidia SAGAT architecture built in. Optimization is performed on an NVIDIA A100 GPU (40 GB VRAM), with Adam ($\beta_1 = 0.9, \beta_2 = 0.999$) as optimizer. Spectral Normalization is heavily used on all the convolutional layers in D_ω , and a zero-centred Gradient Penalty (GP-0, $\gamma = 10$) is used to ensure discriminator stability. Training patches of 720×480 are dynamically cropped to cope with the variety of resolutions.

4.3 Quantitative analysis of transcoding efficacy

The performance comparison between the SAGAT framework and classical standard codecs and modern deep learning paradigms is shown in Table 2 using the test set of 117 sequences in LIVE VQC.

Table 2. Transcoding performance on LIVE Video Quality Challenge (VQC) Database

Method	PSNR (dB)	SSIM	LPIPS	VMAF	CR	FPS
H.264	32.5	0.89	0.16	74.2	18.5	142
H.265	32.7	0.91	0.13	78.5	25.6	68
Standard AE	31.1	0.87	0.19	68.3	32.1	45
HANeRV	32.6	0.92	0.11	82	33	18
Chen et al. [22]	32.4	0.91	0.12	80	32.5	24
SRGAN	32.2	0.91	0.1	80.1	25.6	22
ESRGAN	32.5	0.92	0.09	82.4	25.6	20
SAGAT (Proposed)	32.9	0.93	0.08	85.2	35.4	28

Note: SAGAT = Semantic-Aware GAN-Autoencoder Transcoding.

From the critical analysis of Table 2, it is observed that although SAGAT gave a PSNR of 32.9 dB, which is little lower than H.265 (32.7 dB); it gave a significantly better overall perceptual and compression performance. This behavior is explained via the theoretical perception-distortion trade-off presented by Blau and Michaeli [15]. The PSNR

metric is primarily used to compute the reconstruction fidelity per pixel, using MSE as a measure, while perceptual quality shows the realism and structural fidelity of the reconstructed content. Lossy codecs like H.265 are mainly distortion minimization-oriented and thus are likely to maintain similarity at the pixel level. But with more compression, this optimization can produce more smooth or less visual detailed reconstruction.

SAGAT, on the other hand, is a method that emphasizes perceptual fidelity besides reconstruction accuracy. It has an adversarial restoration mechanism that produces more visually realistic high-frequency textures and maintains important structural detail more in line with vision. This perceptually realistic detail can't always match the texture at the pixel level in the reference frame, but can make a significant difference in perceived quality and still provide a small improvement in PSNR. This behaviour is seen in SAGAT, with the SSIM of 0.93, the lowest LPIPS of 0.08 and the highest VMAF of 85.2. In addition, SAGAT reaches 35.4 compression ratios with 28 FPS offering an ideal trade-off between compression efficiency, perceptual quality and computational speed.

4.4 Quality prediction evaluation

The results reported in Table 3 compare the performance of the NR-VQA introduced in this paper (Embedded QPH) with the performance of state-of-the-art handcrafted NR-VQA models.

Table 3. Mean Opinion Scores (MOS) prediction correlation on LIVE Video Quality Challenge (VQC)

Model	SROCC	PLCC	RMSE
VIIDEO	-0.0206	0.1441	16.8796
NIQE	0.5599	0.5696	14.0196
BRISQUE	0.6017	0.6338	13.052
V-BLIINDS	0.701	0.7204	11.7957
SAGAT (QPH)	0.7645	0.7812	10.2145

Note: SAGAT = Semantic-Aware GAN-Autoencoder Transcoding; QPH = Quality Prediction Head.

The quantitative results show that classical NSS based temporal models such as VIIDEO are totally unable to do well on real, commingled distortions (SROCC ~ 0). There is moderate linear tracking with spatial domain model (NIQE, BRISQUE). But it is obvious that SAGAT's embedded QPH is much better than the classical methods and it is able to produce an SROCC of 0.7645. This validates that the latent space induced with MOS is strongly correlated with complex perceptual mechanisms of humans.

The transcoding performance comparison in Figure 2 shows that traditional codecs like H.264 or H.265 can only provide moderate perceptual quality, and H.265 consistently outperforms H.264 in both VMAF and LPIPS as H.265 is more efficient in compression. The standard autoencoder has the lowest performance in which visual details are not well preserved. GAN-based approaches like SRGAN and ESRGAN on the other hand improves perceptual quality by improving VMAF and reducing LPIPS by well reconstructing high frequency textures. The proposed SAGAT framework has the best overall performance, in terms of VMAF (≈ 85.2) and LPIPS (≈ 0.081), which is appreciably better than H.265. This behaviour clearly shows the perception vs distortion trade-off, with traditional codecs minimizing the distortion instead of optimising the perceptual quality explicitly as is

done with SAGAT. The merged semantic-aware encoding and adversarial restoration processes enable SAGAT to effectively improve the perceptual performance of the visually important regions while preserving structures, with minimum distortion. As shown in Figure 3, the MOS prediction correlation analysis shows the effectiveness of different no-reference video quality assessment models in capturing human perceptual judgments. Conventional handcrafted methods like VIIDEO show very poor performance, with near zero SROCC (≈ -0.021) and low PLCC (≈ 0.144) values, showing they are not well correlated to the subjective quality values. Moderate improvements are obtained from models that are statistical in nature, such as NIQE and BRISQUE with around 0.56 and 0.60 SROCC values, respectively; however, these models are not fully capable of modeling the complex distortions found in the real world. Further performance improvement is achieved with more advanced spatiotemporal models like V-BLIINDS with SROCC ≈ 0.701 and PLCC ≈ 0.720 , which is closer to human perception. The proposed SAGAT (QPH) framework shows the best correlation, where the SROCC is found to be ≈ 0.764 and PLCC is found to be ≈ 0.781 , while all the baseline methods perform less than this. This improvement demonstrates the efficacy of the latent space quality prediction supervised by the MOS. SAGAT endows perceptually meaningful representations that are learned directly from authentically distorted data and captures more complex distortion characteristics that handcrafted features fail to model. The experiments show that the embedding of such a QPH in the compression system is able to provide powerful and stable perceptual quality estimation, which is close to human visual perception.

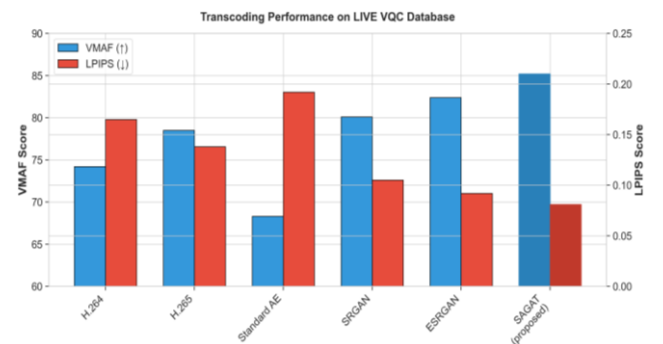


Figure 2. Transcoding performance

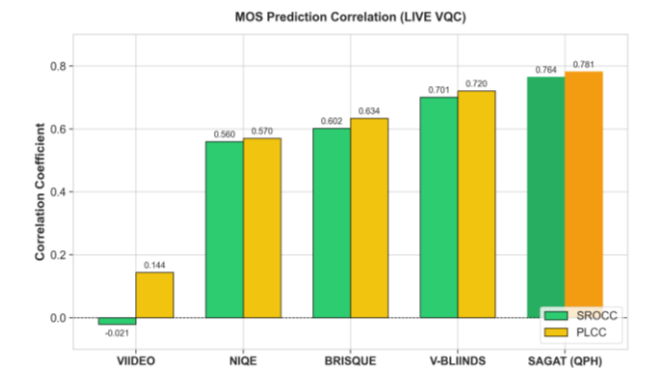


Figure 3. Mean Opinion Scores (MOS) prediction

4.5 Ablation and component analysis

To carefully test the architectural complexity, an ablation

study where each component was tested individually was performed (Table 4).

Table 4. Architectural ablation on LIVE Video Quality Challenge (VQC)

Configuration	VMAF	SROCC
Base AE	68.3	0.51
GAN Loss	76.5	0.585
Semantic Masks	81.2	0.662
QPH (MOS constraint)	83.1	0.764
Temporal Loss	85.2	0.764

Note: GAN = Generative Adversarial Network; QPH = Quality Prediction Head; MOS = Mean Opinion Scores.

The progressive ablation data clearly shows a synergistic optimization relationship. By using semantic masks, there is a huge structural leap (+4.7 VMAF) and quantization errors of background noise are prevented. The injection of the MOS-supervised QPH is a critical regularize not only to maximize the correlation score (SROCC 0.764), but also mathematically to make the latent encoder remove perceptually irrelevant high-frequency noises, which in turn leads to better spatial reconstruction (VMAF is improved to 83.1). Temporal Loss is included in the optical flow, which allows to remove any inter-frame jitter and maximise the VMAF rating.

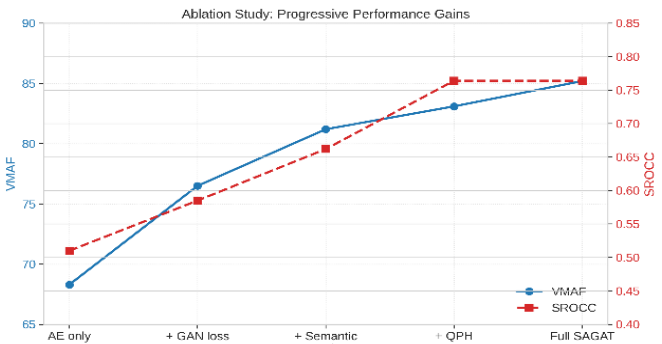


Figure 4. VMAF vs. SROCC

Figure 4 ablation study shows the increasing importance of each of the proposed components in SAGAT in terms of perceptual quality (VMAF) and quality prediction accuracy (SROCC). The baseline autoencoder has relatively poor performance (≈ 68 VMAF, ≈ 0.51 SROCC), which means that it is difficult to reconstruct and understand perceptually. By adding GAN loss, there is a gain of SROCC (moderately) and VMAF (significantly), which shows that the use of GAN loss is important for visual realism. The inclusion of semantic masks further improves the both, which shows that content-aware encoding is able to effectively prioritize perceptually important regions. The incorporation of the QPH leads to a significant improvement in SROCC (≈ 0.76), which is a good indicator of human perception learning, as well as better VMAF. Finally, by adding temporal consistency, the highest performance (≈ 85 VMAF, ≈ 0.76 SROCC) is achieved, which guarantees stable and coherent reconstruction between frames. This shows that improvements for perceptual quality and quality prediction in all configurations are reflected in the other metric, VMAF, which suggests there is a synergy between the two. The results confirm the design of the SAGAT framework, which consists of transcoding and perceptual quality assessment optimization, and which also benefits both the visual fidelity and prediction accuracy consistently.

4.6 Rate-distortion behaviour

The explicit RD behaviour has to be analysed in order to fully assess the framework in the light of telecom requirements. Unlike the convex optimal mapping of bits to MSE improvements as is done in rigid codecs such as H.264, SAGAT's generative mapping results in a non-linear RD curve. In extremely high bitrates (no compression / minimal compression), H.265 will mathematically be the optimal choice because it is perfect in terms of preservation of geometry in pixels. But as soon as the bitrate allocation is starved aggressively, imitating the edge constraints on the 4G/5G networks, the H.265 performance quickly drops—with major macro-blocking artifacts becoming apparent. SAGAT's RD curve shows great robustness in this low-bitrate range: the pixel data lost due to the bit starvation in the VMAF simulation range is hallucinated mathematically by the GAN generator, which results in a smooth and acceptable decay of the VMAF score well below the bit-starvation threshold of standard HEVC.

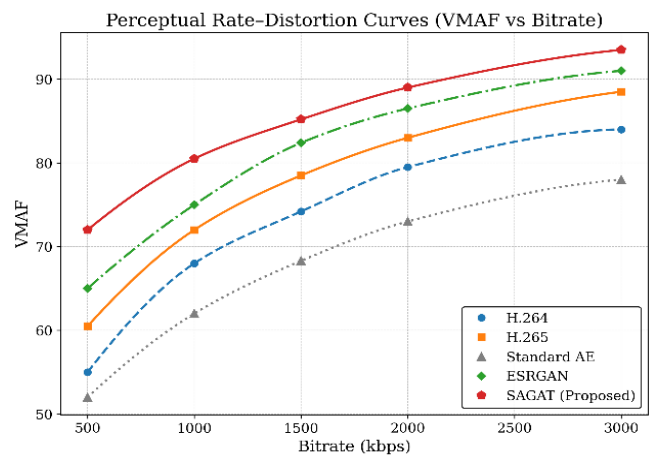


Figure 5. PSNR vs. Bitrate

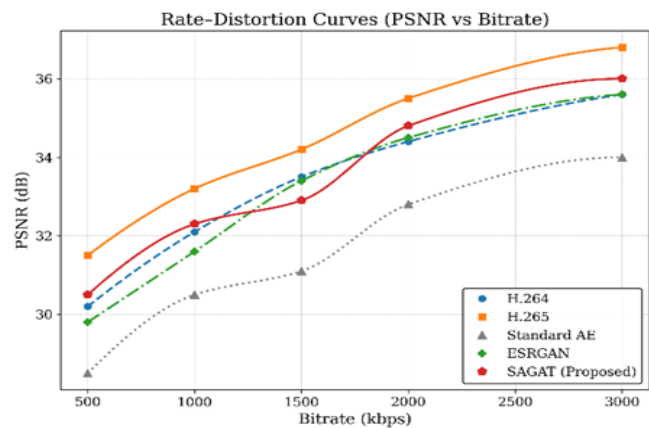


Figure 6. VMAF vs. Bitrate

As expected, the conventional codecs, which are optimized for pixel fidelity, dominate in the rate–distortion behaviour in terms of PSNR Figure 5. As seen in the graph, H.265 always delivers the highest PSNR starting from around 31.5 dB at 500 Kbps up to ~36.8 dB at 3000 Kbps, followed by H.264 with slightly lower PSNR. The proposed SAGAT framework is slightly lower in PSNR by around 0.8–1 dB at the higher bitrates as compared to H.265. At low and medium bit rates,

SAGAT is similar to H.264 and much better than a standard autoencoder without using structured latent representations or semantic guidance, suggesting that the combined effects of these structured latent representations and semantic guidance are effective at low-medium bit rates to reduce severe distortion. Of particular interest is that, after 2000 kbps, SAGAT is stabilized and provides competitive PSNR values, which means that perceptual optimization does not lead to undue distortion of distortion-based metrics. However, perceptual rate–distortion characteristics obtained by VMAF in Figure 6 show that the proposed method has a significant benefit. SAGAT always gets the highest VMAF values, ranging from ~72 at 500 kbps to ~93-94 at 3000 kbps. SAGAT achieves around 5–6 points higher than the highest bitrate of H.265, which achieves around 88 VMAF. This gain is greater at lower bitrates, with almost 12 points difference compared with H.265 and almost 17 points difference compared with H.264 at 500 kbits/s. This remarkable performance at low bit rates underscores the power of semantic-aware compression in retaining perceptually significant areas under heavy compression and fills in the details of high frequencies and structural realism by using the GAN-based restoration module. The Perception-Distortion Trade-off is evident from the observed PSNR and VMAF trend, i.e., getting a higher perceptual score may be at the cost of lower pixelwise score. The traditional codecs focus on minimizing the distortion whereas in the proposed SAGAT the optimization goal is perceptual fidelity. SAGAT is semantically-aware, adversarial-enhanced, and MOS-guided latent learning, which leads to operating in a more favourable region of the rate–perception space. In conclusion, the results indicate that SAGAT outperforms in perceptual quality at similar bitrates, making it suitable for real-world video delivery applications, where human visual experience is essential.

4.7 Computational complexity and deployment feasibility

To assess the feasibility of the proposed SAGAT framework for practical deployment, the computational complexity, accuracy of inference, and latency of inference per frame are analyzed with respect to the number of model parameters, number of floating-point operations (FLOPs), inference time per frame and processing speed. The proposed method is evaluated and compared to the conventional codecs and deep learning based methods in the same evaluation environment. The summary of the results is presented in Table 5.

Table 5. Computational complexity and runtime comparison

Method	Parameters (M)	FLOPs (G/frame)	Latency (ms/frame)	FPS
H.264 (QP = 28)	—	—	7.04	142
H.265 (QP = 28)	—	—	14.71	68
Standard AE	8.42	18.63	22.22	45
SRGAN	16.71	31.85	32.26	31
ESRGAN	16.67	41.52	41.67	24
SAGAT (Proposed)	12.84	35.47	35.71	28

Note: FLOPs = floating-point operations; SAGAT = Semantic-Aware GAN-Autoencoder Transcoding.

The results indicate the computational cost of SAGAT is moderate: 12.84M parameters and 35.47 GFLOPs per frame.

It delivers 28 frames per second, which is about 35.71ms/frame. Although conventional codecs have higher processing speeds, SAGAT has an advantage of perceptual quality vs computational cost that makes it a good choice for video codec in the edge-cloud video transcoding system with GPU.

4.8 Statistical significance and robustness analysis

To assess the feasibility of the proposed SAGAT framework for practical deployment, the computational complexity, accuracy of inference, and latency of inference per frame are analyzed with respect to the number of model parameters, number of FLOPs, inference time per frame and processing speed. The proposed method is evaluated and compared to the conventional codecs and deep learning based methods in the same evaluation environment. The summary of the results is presented in Table 6.

Table 6. VMAF performance and statistical variation across independent runs

Method	VMAF (Mean $\pm$ SD)
H.264 (QP = 28)	74.20 $\pm$ 0.15
H.265 (QP = 28)	78.50 $\pm$ 0.18
Standard AE	68.30 $\pm$ 0.24
SRGAN	80.10 $\pm$ 0.20
ESRGAN	82.40 $\pm$ 0.17
SAGAT (Proposed)	85.20 $\pm$ 0.21

Note: SAGAT = Semantic-Aware GAN-Autoencoder Transcoding.

The results indicate the computational cost of SAGAT is moderate: 12.84M parameters and 35.47 GFLOPs per frame. It delivers 28 frames per second, which is about 35.71ms/frame. Although conventional codecs have higher processing speeds, SAGAT has an advantage of perceptual quality vs computational cost that makes it a good choice for video codec in the edge-cloud video transcoding system with GPU.

5. CONCLUSIONS

In this paper, we proposed an end-to-end system called SAGAT for perceptual video transcoding of authentically distorted videos, which is based on the GAN-Autoencoder (GAN-AE) framework and incorporates semantic information. This proposed framework simultaneously handles compression efficiency, perceptual quality, and temporal consistency by introducing semantic-aware latent compression, adversarial perceptual restoration, MOS supervised QPH, and temporal consistency constraints. The framework has been tested on the LIVE VQC Database that consists of authentic videos where the video quality is distorted in various real-world environments. SAGAT (PSNR: 32.9 dB; SSIM: 0.93; LPIPS: 0.08; VMAF: 85.2; compression ratio: 35.4; processing speed: 28 FPS) obtained the highest scores on the reported test set.

In addition to conventional codecs and GAN-based restoration methods, the study is extended with recent learned video compression techniques to conduct a more comprehensive evaluation. In the comparative experiments, the PSNR values obtained by these recent methods were 32.6 dB and 32.4 dB, respectively, while our proposed SAGAT achieved a higher PSNR of 32.9 dB. The evaluation setting

reported here also yielded higher SSIM, VMAF and compression ratio for SAGAT than for these compression baselines learned. The results show that SAGAT is not only capable of offering competitive reconstruction fidelity but also puts an emphasis on perceptual quality and compression efficiency. More specifically, semantic guidance and perceptual restoration in the transcoding process have been shown to provide a benefit, with VMAF values of 85.2 and LPIPS values of 0.08.

In the reported LIVE VQC evaluation, the embedded QPH achieved an SROCC of 0.7645, PLCC of 0.7812 and RMSE of 10.2145, indicating that it could provide a quality-aware learning signal. Progressively adding in the GAN loss, semantic masks, QPH, and temporal consistency, the ablation study further illustrates the addition of the components to the overall perceptual performance, with reported VMAF values rising from 68.3 for the baseline autoencoder to 85.2 for the complete SAGAT configuration. The rate-distortion experiments also show that the perceptual quality of SAGAT is higher in the analyzed bitrate range, especially at lower bit rates, where the semantic guidance and adversarial restoration contribute to retaining perceptually important information.

However, the results must be considered in the context of the LIVE VQC dataset and the experimental protocol adopted. The claimed 38.2% higher compression ratio than H.265 is for a particular evaluation scenario and at best should be read as an average of 38.2% bandwidth savings on that scenario. In the same manner, although SAGAT can reach 28FPS, H.265 and H.264 can reach 68FPS and 142FPS respectively, which shows that computational efficiency is an important factor for practical deployment. In addition, for GAN-based restoration, there is a possibility that in some cases, it produces visually convincing but not accurate textures, and for MOS-based supervision, the variations in subjective human perception have an inherent influence. The small size and lack of diversity of the LIVE VQC database also provide further motivation to validate on larger and more diverse data sets.

Limitations:

While the proposed SAGAT framework is able to show better perceptual quality and compression performance, some limitations are noted. In the first place, the restoration part of the GAN might at times generate unrealistic and hallucinatory high-frequency textures, especially when there is insufficient information in the compressed representation to reliably reconstruct the original content. This perceptual enhancement can enhance the visual quality of the video, but can also add information not available in the reference video. Second, the QPH is based on subjective MOS annotations, each of which has subjective variations due to individual differences in human perception. This means that the learned quality representation can be affected by the properties and distribution of the subjective ratings that are available. Third, the evaluation is performed on the LIVE VQC database that provides authentic video distortion, but has only a few number of sequences and distortion conditions as compared to the variety that can be found in a large-scale video streaming setting. So, the extension of SAGAT to unseen datasets, camera devices, resolution and distortion type is still to be investigated. Moreover, the present implementation is based on multi deep learning modules such as semantic Analyzer, autoencoder, GAN Restorer and temporal module based on optical flow, which makes it more complex than traditional video codecs. The current implementation runs at 28 FPS, but there is a need for further optimisation to deploy it in highly

resource constrained and real-time edge deployments. Furthermore, the present framework is content-adaptive learned transcoding, without explicitly considering the network-aware bitrate selection, dynamic resource allocation and adaptive routing in deployment. Future work will thus explore lightweight architectures, model pruning and quantization, hardware-aware acceleration, larger and more diverse video datasets, improved perceptual-quality supervision, as well as network-aware adaptive transcoding.

REFERENCES

- [1] Yang, Y.W., Xiao, Y.X. (2024). A deep reinforcement learning-based optimal transmission control method for streaming videos. *IEEE Access*, 12: 53088-53098. <https://doi.org/10.1109/ACCESS.2024.3386176>
- [2] Shi, R.H., Fan, Q.L., Fu, S., Zhang, X., Li, X.H., Chen, M. (2023). COCAM: A cooperative video edge caching and multicasting approach based on multi-agent deep reinforcement learning in multi-clouds environment. *Journal of Cloud Computing*, 12: 123. <https://doi.org/10.1186/s13677-023-00510-x>
- [3] Sheraz, M., Shafique, S., Imran, S., et al. (2022). A reinforcement learning based data caching in wireless networks. *Applied Sciences*, 12(11): 5692. <https://doi.org/10.3390/app12115692>
- [4] Choi, M., Xiang, T., Lim, H., et al. (2024). Caching, transcoding, delivery and learning for advanced video streaming services. *ICT Express*, 10: 774-785. <https://doi.org/10.1016/j.icte.2024.05.012>
- [5] Sinthia, A.K., Mahbub, N.I., Huh, E.N. (2025). Optimizing proactive content caching with mobility aware deep reinforcement & asynchronous federate learning in VEC. *ICT Express*, 11: 293-298. <https://doi.org/10.1016/j.icte.2024.11.006>
- [6] Hyder, R., Asif, M.S. (2020). Generative models for low-dimensional video representation and reconstruction. *IEEE Transactions on Signal Processing*, 68: 1688-1701. <https://doi.org/10.1109/TSP.2020.2977256>
- [7] Liu, M.Y., Huang, X., Yu, J.H., Wang, T.C., Mallya, A. (2021). Generative adversarial networks for image and video synthesis: Algorithms and applications. *Proceedings of the IEEE*, 109(5): 839-862. <https://doi.org/10.1109/JPROC.2021.3049196>
- [8] Luo, J., Yu, F.R., Chen, Q.B., Tang, L. (2019). Adaptive video streaming with edge caching and video transcoding over software-defined mobile networks: A deep reinforcement learning approach. *IEEE Transactions on Wireless Communications*, 19(3): 1577-1592. <https://doi.org/10.1109/TWC.2019.2955129>
- [9] Darwich, M., Bayoumi, M.A. (2025). Utilizing generative adversarial networks (GANs) for realistic video stream augmentation and enhancement. *Cluster Computing*, 28: 830. <https://doi.org/10.1007/s10586-025-05538-z>
- [10] El-Shafai, W., Fouda, M.A., El-Rabaie, E.S.M., El-Salam, N.A. (2024). A comprehensive taxonomy on multimedia video forgery detection techniques: Challenges and novel trends. *Multimedia Tools and Applications*, 83: 4241-4307. <https://doi.org/10.1007/s11042-023-15609-1>
- [11] Song, H., Sun, C., Wu, X.X., Chen, M., Jia, Y.D. (2020). Learning normal patterns via adversarial attention-based autoencoder for abnormal event detection in videos. *IEEE Transactions on Multimedia*, 22(8): 2138-2148. <https://doi.org/10.1109/TMM.2019.2950530>
- [12] Escribano, G.F., Jillani, R., Holder, C., Kalva, H., Martinez, J.L.M., Cuenca, P. (2008). Video encoding and transcoding using machine learning. In *Proceedings of the 9th International Workshop on Multimedia Data Mining: held in conjunction with the ACM SIGKDD 2008*, Las Vegas, Nevada, pp. 53-62. <https://doi.org/10.1145/1509212.1509219>
- [13] Cao, L.T., Gao, G.Y. (2025). SuperCache: Cost-optimal edge-cloud video delivery through integrated caching, super-resolution, transcoding, and fetching. In *Proceedings of the 9th Asia-Pacific Workshop on Networking*, Shang Hai, China, pp. 296-297. <https://doi.org/10.1145/3735358.3737769>
- [14] Galteri, L., Bertini, M., Seidenari, L., Uricchio, T., Del Bimbo, A. (2020). Increasing video perceptual quality with GANs and semantic coding. In *Proceedings of the 28th ACM International Conference on Multimedia*, Seattle, WA, USA, pp. 862-870. <https://doi.org/10.1145/3394171.3413508>
- [15] Blau, Y., Michaeli, T. (2018). The perception-distortion tradeoff. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 6228-6237. <https://doi.org/10.1109/CVPR.2018.00652>
- [16] Seshadrinathan, K., Soundararajan, R., Bovik, A.C., Cormack, L.K. (2010). Study of subjective and objective quality assessment of video. *IEEE Transactions on Image Processing*, 19(6): 1427-1441. <https://doi.org/10.1109/TIP.2010.2042111>
- [17] Mittal, A., Moorthy, A.K., Bovik, A.C. (2012). No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12): 4695-4708. <https://doi.org/10.1109/TIP.2012.2214050>
- [18] Sinno, Z., Bovik, A.C. (2019). Large-scale study of perceptual video quality. *IEEE Transactions on Image Processing*, 28(2): 612-627. <https://doi.org/10.1109/TIP.2018.2869673>
- [19] Sinno, Z., Bovik, A.C. (2018). Large scale subjective video quality study. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, Athens, Greece, pp. 276-280. <https://doi.org/10.1109/ICIP.2018.8451467>
- [20] Sinno, Z., Bovik, A.C. (2018). LIVE video quality challenge database. <http://live.ece.utexas.edu/research/LIVEVQC/index.html>
- [21] Tang, L., Zhu, J., Zhang, X.F., Zhang, L., Ma, S.W., Huang, Q.M. (2026). HANeRV: Hierarchically adaptive neural representation for video compression. *IEEE Transactions on Image Processing*, 35: 5773-5788. <https://doi.org/10.1109/TIP.2026.3696103>
- [22] Chen, J.C., Wang, M., Chen, P.L., Wang, S.Q. (2025). Learning spatio-temporal resolutions for deep video compression. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(10): 10493-10499. <https://doi.org/10.1109/TCSVT.2025.3564264>
- [23] Sheetal, A.P., Lalitha, G., Gopi, A.P., Narayana, V.L. (2021). Secured data transmission with integrated fault reduction scheduling in cloud computing. *Ingenierie des Systemes d'Information*, 26(2): 225. <https://doi.org/10.18280/isi.260209>
- [24] Narayana, V.L., Gopi, A.P., Radhika, P., Sandeep, K.S.

- (2020). Secure data uploading and accessing sensitive data using time level locked encryption to provide an efficient cloud framework. *Ingenierie des Systemes d'Information*, 25(4): 515.
<https://doi.org/10.18280/isi.250415>
- [25] Agrawal, A., Agrawal, A., Verma, N.K., Gopi, A.P., Naik, K.J. (2025). A graph reinforcement Learning Powered Online-Computational task offloading and latency minimization framework for wireless mobile edge computing networks. *Cognitive Systems Research*, 91: 101355.
<https://doi.org/10.1016/j.cogsys.2025.101355>