IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 13 October 2024, accepted 29 October 2024, date of publication 4 November 2024, date of current version 26 November 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3491613

SURVEY

A Survey on QoE Management Schemes for HTTP Adaptive Video Streaming: Challenges, Solutions, and Opportunities

REZA KALAN[1] **AND ISMAIL DULGER**[2]
[1]Department of Computer Engineering, Istinye University, 34396 Istanbul, Türkiye
[2]Digiturk beIN Media Group, 34320 Istanbul, Türkiye

Corresponding author: Reza Kalan (reza.kalan@istinye.edu.tr)

This work was supported by Digiturk beIN Media Group's Research and Development Team (https://www.digiturk.com.tr/) solely for funding the APC, in close cooperation with Istinye University (https://www.istinye.edu.tr/).

ABSTRACT With the emergence of new video streaming technologies, advancements in networking paradigms, and the increasing popularity of mobile and smart devices, we are witnessing phenomenal growth in live video traffic over the Internet. To effectively address the explosive growth of multimedia applications over the Internet, it is crucial to consider scalability, quality, and security. The reliability of HTTP Adaptive Streaming (HAS), which leverages TCP, encourages many Over-the-Top (OTT) providers to adopt progressive streaming technology. Monitoring network traffic patterns and client behaviors provides client-side players with greater intelligence to adapt suitable video quality. However, the inflexibility and inefficiency of legacy networks and streaming applications often diminish the perceived streaming quality for clients. This study aims to explore the interplay between machine learning, emerging network architectures, and streaming technology paradigms. Furthermore, it survey the technical challenges within the adaptive video streaming and content delivery technologies, where adaptive streaming leverages advancements in network and artificial intelligence paradigms, edge computing, and NFV-SDN technologies to better adapt to network dynamics and enhance Quality of Experience (QoE). This study focuses exclusively on papers published in the last five years.

INDEX TERMS Adaptive video streaming, video codec, multimedia system, edge computing, QoE, CDN, NFV-SDN.

I. INTRODUCTION

Content delivery over the Internet in the form of streaming is a technology used to allow users to display video or listen to audio almost immediately, even before the entire file is completely downloaded. A relatively fast Internet connection is required to provide streaming services at sufficient quality and to avoid re-buffering for the receivers. Streaming content, whether live or on-demand, uses a significant portion of IP network traffic. On a global scale, over 70% of the global population will have mobile connectivity, 66% of the global population (5.3 billion) will be Internet users by 2023, and video will make up 82% of all IP traffic. In addition, 66% of connected flat-panel TV sets will be 4K [1]. However, current streaming solutions struggle to simultaneously address the demands for scalability, quality, and security amid the rapid growth of multimedia applications. The growing popularity of smart devices and ubiquitous multimedia services is driving greater streaming content usage, which in turn elevates network resource consumption and directly impacts perceived video quality.

Adaptive Bitrate (ABR) streaming technology refers to methods of delivering video content over the Internet that adjust to dynamic network conditions and device capabilities. It works by adaptively selecting the appropriate video bitrate based on network throughput, buffer occupancy, and device performance. HTTP Adaptive Streaming (HAS) technologies include Apple's HTTP Live Streaming (HLS) [2], Microsoft

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

Smooth Streaming (MSS) [3], and Dynamic Adaptive Streaming over HTTP (DASH) [4]. These technologies aim to fill players buffer's with appropriate content in the form of bitrate-they need next. Bitrate which indicates the amount of data encoded per second is affected by network and player conditions which change from time to time. This adaptivity in seamlessly switching between different quality levels and minimizing buffering ensure a smoother video playback experience.

Delivering low-latency streaming content to end-user over heterogeneous networks where the traffic pattern changes rapidly is a major challenge. The pivotal aspect of content delivery lies in network infrastructure, particularly during peak traffic periods. The most common streaming problem relates to the availability of network bandwidth. As shown in Fig. 1, insufficient available bandwidth can negatively impact media delivery. As a result, media player on shared bandwidth, may experience poor quality or re-buffering. A different class of solutions has been discussed to address this problem. Storing content closer to the end-user reduces latency and improves response time, but it is costly. Leveraging Content Delivery Networks (CDN) and distributed edge servers across wide geographical regions can effectively address the bottleneck issue while streaming from a single server. The unstable nature of wireless channels and the characteristics of fixed and mobile networks in heterogeneous environments are also significant concerns. Delivering low-latency multimedia content to end-users across heterogeneous networks, characterized by rapidly changing traffic patterns, presents a considerable challenge [5].

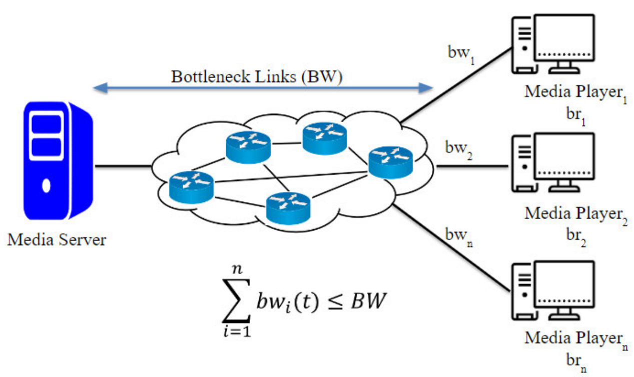


FIGURE 1. Fluctuations in bandwidth and bottleneck link issues can impact end-user perceived QoE.

Despite the flexibility that CDN and adaptive video streaming provides, legacy network architecture based on the pure Transmission Control Protocol/Internet Protocol *(TCP/IP)* is no longer efficient. A clean-state approach to completely redesigning the current network is more costly and practically impossible. Network configuration is another concern that must be considered. The heterogeneous nature of technologies, infrastructure, and applications makes managing such a system challenging. While network conditions are often ad hoc, configuring the network based on predicted policies can be difficult. In addition, any alterations in policy or faults will require reconfiguration of network devices.

To address this problem, we need network transformation. The results of the most agile, flexible, and programmable network infrastructures are better aligned with the needs of application workloads. Multi-access Edge Computing (MEC) [6], [7], [8], network slicing, Network Function Virtualization (NFV) [9], [10], [11], Software Defined Networking (SDN) [12], [13], [14], and Machine Learning (ML) [15], [16] algorithms and technologies aim to overcome the limitations of the current network design. Edge computing significantly reduces latency and allows the maintenance of a fast and reliable connection by offloading traffic. The NFV is an abstract of network virtualization that addresses various networking challenges without touching the underlying infrastructures. Moreover, decoupling network functions and the control plane from the data plane enhances the intelligence of central management and the agility of infrastructure forwarding devices. Such an agile and flexible environment meets network scales and applications' changing demands with a seamless user experience. In recent years, the reliability and adoption of IP services have driven the demand for high-quality multimedia systems, prompting both media providers and telecom operators to upgrade their systems and transition to advanced network technologies. In this evolution, redefining network architecture and employing intelligent solutions are crucial.

A. SURVEY NOVELTY AND SCOPE OF CONTRIBUTION

The main objective of this study is to share a comprehensive overview of QoE enhancement in current and future video delivery systems. This study aims to survey adaptive streaming technology and investigate factors that can impact video content delivery, improving performance in terms of the network's Quality of Service (QoS) and clients' Quality of Experience (QoE). Table 1 lists different previous survey studies that attempt to address QoS/QoE management in video applications by identifying different approaches, use cases, and architectures. Most existing surveys only addressed the quality challenges (e.g., re-buffering, bitrate, and stability) of multimedia streaming services in specific domains.

Pretty old but popular work is the study of Barakovi et al. [17] who presented a survey on QoE management in wireless networks without considering HAS-QoE and potential future network technologies. Seufert et al. [18] provided an overview of the technical factors affecting HAS-QoE, but did not discuss potential future technologies. Zhao et al. [19] presented a survey on QoE in video transmission. Su et al. [20] present a survey on QoE management in video streaming in wireless networks. Similar to previous studies, these works also lack a discussion of future networks. Skorin-Kapov et al. [21] survey state-of-the-art findings and present emerging concepts and challenges related to managing

QoE for current and future networked multimedia services. Benmir et al. [22] present the most important factors affecting end-user's QoE during video streaming over wireless, ad-hoc, and mobile networks, with a special focus on VANET.

Petrangeli et al. [23], Bentaleb et al. [24], Barakabitze et al. [25] attempt to address how programmability and flexibility of future network technology, such as NFV-SDN can optimize media service delivery and improve HAS-QoE. Barman et al. [26] provide an overview of HAS-QoE, classifying existing models and analysing factors that affect video quality without delving deeply into emerging network technology. Sousa et al. [27] presented a basic concept of QoE-based wireless resource scheduling. However, this study also has no records of the use of NFV-SDN, edge, or cloud technologies. Barakabitze et al. [28] discuss the advantages of leveraging NFV-SDN for network slicing in 5G networks.

Yaqoob et al. [29] survey and explores adaptive 360° video, presenting an end-to-end delivery solution while demonstrating ongoing technological advancements and standardization efforts in this area. Siriwardhana et al. [30] provide a comprehensive overview of mobile augmented reality in 5G networks, focusing on architectures, applications, and technical aspects of MEC. Jiang et al. [31] surveyed the taxonomy of MEC-enabled video streaming applications from the perspectives of intelligent video acceleration, video streaming analysis, augmented reality services, and connected vehicles. Kougioumtzidis et al. [32] focused on optimizing mobile multimedia streaming service delivery by evaluating and predicting end-user QoE, with a focus on extended reality and video gaming applications.

B. CONTRIBUTION BEYOND PREVIOUS SURVEYS

Several survey papers have been published in recent years providing various overviews of studies aimed at improving multimedia QoE. A subset of these studies focuses on the challenges of QoE management in multimedia systems and QoE-driven network management approaches [21], as well as on QoE assessment through coordinated application-network interaction [32]. Unlike previous studies, this research introduces a pioneering concept of a QoE management scheme specifically for HAS technology, emphasizing comprehensive end-to-end solutions, as well as the associated opportunities and challenges. To address this gap, the study focuses exclusively on the significant benefits that intelligent machine learning and network-assisted approaches offer for QoE management in HAS technology and its applications. The main contributions of this study are outlined below.

- We will begin by discussing the QoE monitoring and management parameters that affect the quality of video streaming services. Next we will present the difference between QoE and QoS. In addition, we provide an overview of virtualization and network software in the future network.
- A comprehensive overview of video streaming technologies will be provided, covering both Non-HAS and HAS-based technologies, as well as aspects such as video coding, video packaging, player models, and adaptation schemes.
- Deal with deep details of existing bitrate adaptation schema in HAS technology and discuss the advantages and disadvantage of server-side and client-side adaptation algorithms.
- Follow the direction and movement of HAS technology and improving QoE within existing solutions. We will outline the video and network requirements for both Video on Demand (VoD) and live applications.
- We will discuss the challenges and future research directions related to network-assisted and intelligent adaptive streaming.

C. STRUCTURE AND ORGANIZATION

To this end, this paper presents a comprehensive review and thorough analysis of recent research on HAS-QoE, focusing on intelligent and network-assisted approaches. Most of the articles reviewed in this study were published after 2020 and are generally not covered by the existing survey articles. We also include some previous works published before 2020 to provide a broader context for the development of this research area. The rest of this survey study is organized as follows:

- *Section II*: This section provides an overview of QoE, including its influencing factors, monitoring, management, and optimization. It also describes the relationship between QoE and HAS schemas, focusing on aspects such as stability, fairness, and optimal network resource utilization.
- *Section III*: This section briefly discusses video coding and compression techniques. It also classifies various video streaming technologies and explains the advantages of adaptive streaming over legacy streaming methods.
- *Section IV*: This section covers HTTP adaptive video streaming technology and terminology. It discusses bitrate adaptation schemes and the decision points for adaptation. Finally, it provides a detailed overview of the technology and the relationships among various stakeholders in the video content delivery chain, including content distributors, Internet Service Providers (ISPs), and CDNs.
- *Section V*: This section examines optimization techniques for origin media servers and encoding methods to enhance QoE, taking into account network constraints and users' resource limitations. Furthermore, it discusses the pros and cons of each proposed solution. Also, the survey provides an overview of security challenges in digital content distribution, including the implementation of Digital Rights Management (DRM) and watermarking models, which can significantly affect the end-users' QoE.

TABLE 1. A summary of related survey papers and covered topics.

Survey Papers	Year	Topics Covered and Scope	Domain	Future Technologies
Barakovi *et al.* [17]	2013	Challenge of QoE management in wireless networks	Wireless-QoE	✗
Seufert *et al.* [18]	2015	Present technical influence factors affect HAS-QoE	HAS-QoE	✗
Zhao *et al.* [19]	2016	Survey QoE in video transmission	General view	✗
Su *et al.* [20]	2016	Survey QoE in video streaming over wireless networks	Wireless networks	✗
Skorin-Kapov *et al.* [21]	2017	Survey concepts and challenges in QoE Management	HAS-QoE	Partially
Benmir *et al.* [22]	2018	Survey QoS/QoE in vehicular ad-hoc networks (VANETs)	VANET-QoE	✗
Petrangeli *et al.* [23]	2018	Survey QoE-Centric adaptive streaming status & challenges	HAS-QoE	SDN only
Bentaleb *et al.* [24]	2019	Survey state-of-the-art bitrate adaptation algorithms	HAS-QoE	NFV-SDN
Barakabitze *et al.* [25]	2019	QoE management of multimedia services in future networks	HAS-QoE	NFV-SDN, Edge, Cloud
Barman *et al.* [26]	2019	Survey QoE modeling for adaptive streaming technology	HAS-QoE	✗
Sousa *et al.* [27]	2020	QoE-oriented wireless resources scheduling	Wireless sensor	✗
Barakabitze *et al.* [28]	2020	Survey network slicing using NFV-SDN	5G network	NFV-SDN
Yaqoob *et al.* [29]	2020	Survey on adaptive 360° video streaming	Virtual reality	360° video
Siriwardhana *et al.* [30]	2021	Survey on mobile augmented reality with 5G	5G network	MEC, 5G
Jiang *et al.* [31]	2021	Multi-access Edge Computing applied to video streaming	Video streaming	MEC, NFV-SDN
Kougioumtzidis *et al.* [32]	2022	Survey on ML-based QoE prediction models	Video gaming	Machine learning
		i) A tutorial on adaptive video streaming technologies	Deep overview	MEC
		ii) A tutorial on modern network architecture and design	NFV-SDN/Edge	Edge, Cloud/Fog
Our work		iii) Survey QoE management in legacy & modern networks	HAS-QoE	NFV-SDN, 5G
		iv) Survey QoE management with intelligent approaches	HAS-QoE	Machine learning
		v) Benefits & challenges of implementing security	HAS-QoE	DRM, Watermark, QUIC

- *Section VI*: This section covers emerging applications and energy-aware adaptation technologies aimed at enhancing QoE. It also includes an overview of video security and the challenges related to the illegal redistribution of video content. Additionally, it discusses the use of Artificial Intelligent (AI) in video streaming and outlines potential future directions in this area.
- *Section VII*: This section classifies cutting-edge network technologies aligned with media streaming and explores the challenges associated with various streaming use cases (e.g., live, VoD, etc.).
- *Section VIII*: This section survey on virtualization and softwarization technologies, exploring the correlation between NFV and SDN. It discusses how leveraging the flexibility and programmability of NFV-SDN can enhance perceived QoE.
- *Section IX*: This section conclude the paper and draw future path in this direction.

II. OVERVIEW ON QOE MANAGEMENT AND VIDEO STREAMING

In this section, we have a quick but accurate overview of the definitions, concept, and objectives of QoE. We also highlight Key Performance Indicators (KPIs) and metrics that influence QoE. Finally, we discuss the history of video streaming technology with focusing on the conventional streaming protocol (Non-HAS-based) and HTTP-based adaptive (HAS-based) protocol.

A. VIDEO QUALITY AND QOE PERSPECTIVE

The objective of the multimedia system, which includes capture, transfer, and display, is to deliver high-quality services to end-users while minimizing latency and re-buffering. From the client's perspective, QoE is *"the degree of delight or annoyance of the user of an application or service"* [33]. From the service provider's perspective, QoE is crucial for maintaining client satisfaction and reducing churn [34]. QoE-driven approaches aim to maintain quality above a certain threshold in the shared resources. However, QoE fairness does not always refer to equal quality. QoE management aims to maximize the overall quality of service across multiple clients (devices or players) while achieving an optimal level of QoE. A critical aspect of this process is analyzing the provided services in terms of QoS/QoE parameters.

1) QOS AND QOE DEFINITION

QoS is a set of technologies operating at the network and infrastructure levels to reliably support high-priority applications and traffic. It refers to any technology that manages data traffic to reduce packet loss, latency, and jitter in the network during data transmission. While QoS is the description or measurement of the overall performance of a service, QoE is a measure of the overall level of client satisfaction and experience of a product or service. The primary goal of QoE management is to optimize end-user experience while efficiently using network resources. Accurate QoE prediction requires high consistency with the Human Visual System (HVS), low latency, and unbiased evaluation, which are difficult to simultaneously realize in practice [35].

Effective QoE management for a specific application involves understanding and identifying a set of influencing factors, both subjective and objective, under the supervision of multiple components within the service supply chain [32], [36]. QoE is a subjective index that correlates customer perceptions, expectations, and experience with application and network performance. In other words, QoE considers the user's subjective perception and expectations of a given service [37]. Ultimately, we can encapsulate QoS within the broader framework of QoE.

2) QoE MODELS AND INFLUENCE FACTORS

Delivering low-latency streaming content to end-users presents a significant challenge, particularly across diverse networks with rapidly changing traffic patterns. The availability and optimization of network resources are critical to ensuring a seamless streaming experience. Achieving a higher QoE during streaming is complex task due to several factors, including various connected client types, client distribution, request and traffic patterns, diversity in video formats and bitrates, as well as fluctuations in ISP and CDN performance. Beyond the impact of the QoS and QoE relationship on the quality of multimedia delivery, it is important to consider a set of human-centric metrics. Fig. 2, summarizes three main factors that influence QoE: system influencing, context influencing, and human influencing factors.

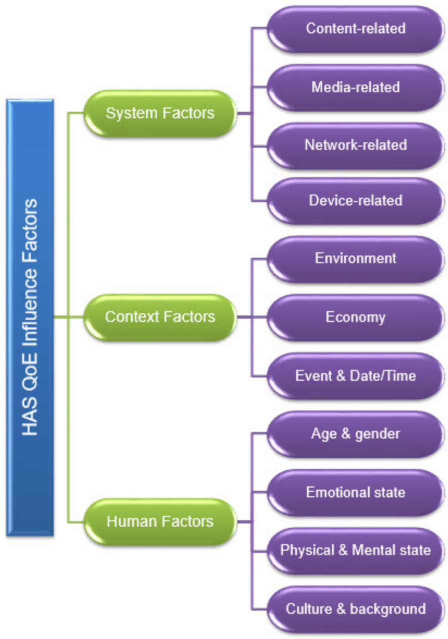


FIGURE 2. **Factors Influencing perceived video quality.**

3) QoE MONITORING, MANAGEMENT AND OPTIMIZATION

Managing and optimizing the perceived quality for end-users involves monitoring the entire glass-to-glass media delivery ecosystem. This observation and control enable the detection and optimization of the root causes that may lead to quality degradation. Throughout the media delivery workflow, video quality can deteriorate at any stage, directly impacting QoE:

- *Capturing*: Describe the process of converting source video into digital format. The resulting digital data is referred to as a video stream.
- *Encoding*: This process involves creating multiple copies of a single file, each with a different bitrate. Video codecs compress video size, typically using lossy compression, which means that some information from the original video is lost in the compression process.
- *Packaging*: The encoder ingests video files into the origin server, which is responsible for video packaging. Packaging involve converting the original compressed video content into different video formats.
- *Distribution*: Distribution of video content to a wide geography through CDN networks. This process brings the content as close as possible to the end-users, ensuring efficient delivery and reduced latency.
- *Displaying*: Various types of video players are used for visualizing video content. In the business model, those device refers to clients (e.g., smart TV, mobile, web, etc).

4) QoE AND HAS SCHEMA

The ability to measure QoE provides OTT with insight into network performance metrics such as reliability, availability, scalability, speed, and efficiency, all are crucial for enhancing overall client satisfaction. The most common mechanisms used to improve the quality of end-users are either based on network optimization (network-assisted resource allocation) or client-based adaptive video streaming. Recent research increasingly emphasizes intelligent and network-assisted approaches that integrate network decisions to enhance quality of service and ensure high user satisfaction by optimizing network resources, as discussed in this study. Research has demonstrated that employing a centralized management controller can significantly improve end-user overall video quality [18]. In this context, Bentaleb et al. [24] have argued that a robust HAS scheme should achieve three main objectives:

- *Stability*: Refers to quality oscillations, which negatively affect video quality. Clients should avoid frequent bitrate switching.
- *Fairness*: This involves sharing bandwidth among HAS clients, considering content, clients, and network characteristics.
- *High Utilization*: Refers to optimizing network resource usage to achieve high performance without compromising client stability and fairness.

B. VIDEO QUALITY AND QUALITY KPI

QoE assessment can be conducted through two primary methods: subjective and objective evaluation. Subjective assessment methodologies involve gathering feedback from human assessors who undergo various tests or are exposed to stimuli. In contrast, objective assessment models evaluate QoE solely based on objective quality metrics.

1) SUBJECTIVE QoE MANAGEMENT

Since the human visual system is the ultimate receiver of adaptive video streaming, subjective evaluation remains the simplest and most reliable approach to assess ABR techniques [38]. The KPI refer to parameters that influence video quality. Optimizing the playback experience involves seamless video playback, quick startup, consistently high-quality, and stable images without buffering. Based on our empirical observations of real-world video streaming, the key factors to measure video streaming performance are:

- *Video Bitrate*: Each video file is encoded at a multiple bitrates, known as a representation, which indicate the number of video bits that can be transmitted in a certain amount of time. Higher bitrates are associated with better video quality. End-users prefer to play a video with higher quality, thus popular video encoded with higher bitrate. However, achieving a higher bitrate requires longer encoding and packaging time on the video provider's side and demands greater bandwidth for distribution. Additionally, end-users require a high-speed network connection and a robust display device to fully enjoy the higher-quality video.
- *Startup Time*: This refers to the duration spent filling the buffer before the video initiates. Displaying video immediately after completing the download segment results in fast re-buffering. Thus, there is a threshold for video buffering before playback can begin. Conversely, larger threshold values lead to extended startup times. Our statistics (in real-world streaming) indicate that users tend to abandon waiting for a video if its startup exceeds a few seconds.
- *Lag Length*: When the buffer reaches the threshold level, the video starts. During display, if the download rate can keep pace with the bitrate, the viewer will experience smooth playback. However, this is not always the case. Sometimes, the buffer is drained, and the video stops because there isn't enough video to continue. For example, utilizing high bitrate settings can put a significant load on your system, potentially leading to lagging issues.
- *Quality Oscillation*: The player adaptively switches between different bitrates to avoid buffering issues. However, frequent oscillations between bitrates disrupt the seamless stream and diminish user satisfaction due to poor video quality.

2) OBJECTIVE QoE MANAGEMENT

Objective models assess subjective quality solely through objective measurements or indices. Their advantages include ease of implementation and modification, focusing on measurable QoS factors and mathematical models. However, a drawback of the objective approach is its potential inaccuracy, providing only an approximation rather than a precise measure of end-user perceived quality [32]. Over the years, researchers have dedicated significant effort to developing metrics and models that can objectively predict the quality of multimedia services as perceived by end-users. Table 2 summarized some objective quality assessment metrics applied in video applications. Objective models for assessing subjective quality are classified into three categories [32], [39]:

- *Full Reference(FR)*: Assessing the quality of a test image by comparing it with a reference image that is assumed to have perfect quality.
- *Reduced Reference(RF)*: Assessing the quality of a test and reference image based on a comparison of features extracted from both images.
- *No Reference(NR)*: Assessing the quality of a test image without any reference to the original image. This method relies on image evaluation based on internal features and statistical models without comparing it with an external reference image.

III. MULTIMEDIA SYSTEM AND STREAMING TECHNOLOGIES

Before delving into streaming protocols and describing the older delivery methods that preceded emerging HAS, it is beneficial to provide a brief overview of video coding technology.

A. VIDEO CODING AND COMPRESSION

Video coding technology is utilized across various applications, ranging from social media, video conferencing, wireless and Internet video streaming, to standard and high-definition TV broadcasting. Delivering video content to a range of decoding devices with varying computational power and displays diversity, requires flexible adaptation to optimize quality under limited network resources and conditions. In the QoE concept, the effectiveness of video coding standards is crucial for streaming video. The key point is that all streaming solutions use advantage of a compression technology known as encryption to maximize bandwidth usage and enable real-time retrieval. From a content perspective, QoE is closely tied to the quality of the encoded video. Advanced Video Coding (AVC/H.264) [40] is currently the most supported video codec by the existing streaming platforms. However, new generation video codecs such as High-Efficiency Video Coding (HEVC/H.265) [41], Video Processor 9 (VP9) [42], AOMedia Video 1 (AV1) [43], and Versatile Video Coding (VVC/H.266) [44] have been developed due to their higher compression efficiency. Table 3 shows current popular video codecs used by most OTT. Some of these are licensed and some are free of charge.

1) ADVANCED VIDEO CODING

In the last two decades, several specific scalable video profiles have been incorporated into video codecs such as MPEG-2, H.263, and MPEG-4. However, all these solutions present a reduced coding efficiency when compared with non-scalable video profiles. As a consequence, scalable profiles have been scarcely utilized in real applications [45]. AVC/H.264, also referred to as MPEG-4 part 10, is a video compression standard that can support up to 4K resolution and is widely used for recording, compressing, and video con-

TABLE 2. Well-known objective quality assessment metrics applied in video applications.

Methods	Abbreviation	Description	Ref.
PSNR	Peak Signal-to-Noise Ratio	Ratio between the maximum power of a signal and the power of distorting noise	FR
VSNR	Visual Signal-to-Noise Ratio	Analyze how pixel-binning and resizing methods influence noise visibility	FR
VMAF	Video Multimethod Assessment Fusion	Predicts subjective video quality based on a reference and distorted video sequence	FR
VQM	Video Quality Metric	Measures perceptual effects of video impairments (e.g., blurring, color distortion)	NR
MOVIE	MOtion-based Video Integrity Evaluation	Space-time domain of evaluation special distortion, by motion tracking in video	FR
MPQM	Moving Pictures Quality Metric	A channel-based distortion measure, accounting for contrast sensitivity and masking	FR
MS-SSIM	Multi-Scale Structural SIMilarity	Measuring the similarity between two images	FR

TABLE 3. Popular video codec used for coding and compressing video content.

Codec	Name	Container	HDR	VFR	Compression Method	RTP, WebRTC
AV1	AOMedia Video 1	ISOBMFF, MP4, MPEG-TS, WebM	✓	✓	DCT	✓
VP9	Video Processor 9	3GP, Ogg, WebM	✗	✓	DCT	✓
AVC/H.264	Advanced Video Coding	3GP, MP4, MPEG-TS	✓	✓	DCT	✓
HEVC/H.265	High Efficiency Video Coding	ISOBMFF, MP4,MPEG-TS	✓	✓	DCT	✗

tent distribution. The intent of the AVC/H.264 definition was to create a standard capable of providing good video quality at much lower bitrates than previous standards (such as MPEG-2, H.263) without increasing design complexity and provide enough flexibility to be applied to a wide variety of applications on a heterogeneous network. Almost all general coding standards such as H.261, H.263, and AVC/H.264 have the same basic functionality, but AVC/H.264 achieves better efficiency in terms of compression. Today's AVC/H.264, as a popular video coding standard, is widely used in video streaming systems. However, the maximum resolution that AVC/H.264 can support is 4K (4096×2160), which is not sufficient for next-generation streaming such as virtual reality or 8K video.

It is worth reminding that there are two important video standardization groups. i) Video Coding Experts Group (VCEG) under the International Telecommunication Union-Telecommunication Standardization Sector (ITU-T). This group is responsible for standardizing H.26x family coding standards, and ii) Moving Picture Experts Group (MPEG) from the International Organization for Standardization (ISO) which developed MPEG-1, and MPEG-4 standards. These two groups jointly cooperated and introduced H.262/MPEG-2, H.264/MPEG-4 advanced video coding, and HEVC/H.265 standards [41]. It has been proved that in the same video quality, HEVC/H.265 achieves 50% better bitrate reeducation compared with AVC/H.264. However, it is not fully supported by all players. Based on our experiment in a real video streaming platform, legacy devices have difficulty meeting HEVC/H.265 requirements.

2) SCALABLE VIDEO CODING

Scalable Video Coding (SVC), which uses AVC/H.264 standard, is a technology that encodes content in dependency layered format. In terms of quality, *scalability* indicates that a video decoder can retrieve video sequences without receiving all quality layers [46]. In the SVC technology, several sub-streams with different qualities are transferred in one stream. As a rule, these streams are basic and enhancement ones. While the first layer provides the base quality, enhanced layers increase the display quality by adding additional data with respect to improving bitrate. In other words, SVC introduces a way to transfer additional streams within a stream. Secure transmission of the base stream across the network is essential for SVC even though the enhancement layers are subject to packet loss. This may require additional processing at the endpoint to support continuous packet loss monitoring.

Layer dependency forces players to receive layers (or sub-streams) in an orderly fashion. While weaker players have the lowest bandwidth request and display base layer, players with high enough bandwidth can experience better quality by requesting base layer and a few enhanced layers. For example, to display a video with two quality layers, the player must request the base and the first enhancement layers. This unique encoding scheme enables the quality of video segments to change incrementally, leading to greater flexibility and better adaptability in high dynamic networks [47]. Recall that, generating VBR-encoded segments is easy, but not easy to stream.

3) MULTIPLE DESCRIPTION CODING

Like the SVC standard, Multiple Description Coding (MDC) encodes the video into layers called descriptions [48]. It is an effective tool against burst packet losses in error-prone networks. MDC uses advanced video coding tools and features provided in AVC/H.264 to introduce redundancy in descriptions [49]. Similar to SVC, each description received by the client improves the video quality. However, in contrast to SVC, the description layers are self-decodeable and each layer carries basic information to decode the video independently. This characteristic of MDC causes extra overhead compared to SVC as each description contains redundant information to be decoded without requiring other descriptions. Thus, MDC is more tolerant of loss compared to SVC, but SVC provides better compression.

To illustrate the effects of packet loss at the application layer and the received quality, we provide an example with

TABLE 4. Example of SVC and MDC codec behaviors.

	Received Representations				Displayed Quality	
Representation	$\#R_1$	$\#R_2$	$\#R_3$	$\#R_4$	SVC	MDC
Scenario 1	✓	✗	✓	✗	1	2
Scenario 2	✓	✓	✗	✓	2	3

a multi-layer codec video. Suppose a video with one base layer and three enhancement representations is encoded using SVC. The same video is alternately encoded with four representations using MDC. For instance in the SVC codec, to achieve a quality equivalent to three layers, the base layer and the first two enhancement representations (i.e. R_2, R_3) should be received by the client. On the other hand, in the MDC codec client just needs to receive any three layers to play the video with the same quality. Table 4 demonstrates the received video quality for the given scenarios for SVC and MDC clients. In the table, R_1 indicates the base layer for the SVC codec. For both scenarios, the client streams video at a higher quality with the MDC codec than the quality received with the SVC codec.

B. VIDEO STREAMING TECHNOLOGIES

1) PROGRESSIVE AND ADAPTIVE STREAMING

Users want to display video files in the best quality without interruption and wasting time on downloads. Modern video streaming addresses this problem with two different technologies:

- *Progressive Streaming*: Progressive streaming involves transferring video files from a server to a client, typically using the HTTP protocol. Unlike regular downloads, the client may start playing the video before completing a download. However, on a poor network connection where video stream cannot be processed quickly enough, a progressive video will pause.
- *Adaptive Streaming*: Adaptive streaming involves encoding multiple live or on-demand streams and dynamically switching between them based on the current network bandwidth or client resources. Unlike download mode, which requires the entire video file to be downloaded before playback, streaming technology allows a video to be played as parts of it are received. In this framework, live and VoD streaming represent the two primary types of content delivery.
  - Video on Demand: In on-demand distribution, users have the freedom to choose and play any video at any time. This flexibility increases distribution cost (e.g., origin server load, network bandwidth), as each user requires an individual network stream.
  - Live Streaming: In live streaming, multiple users simultaneously access and display the same video content. Therefore, pushing and caching content at the edge points reduces the origin server's processing cost and improves network bandwidth utilization.

Also, it is good to have a brief overview of the two main transport layer protocols: TCP and User Datagram Protocol (UDP), which differ in speed, quality, and reliability. While TCP is quite reliable, UDP is agile and fast. In addition, TCP is the unicast protocol used for communication between two endpoints, whereas UDP can handle unicast, multicast, and broadcast communications. Preferring the right transport layer protocol depends on application specifications. For example, HTTP-based streaming uses TCP protocol. Quick UDP Internet Connections (QUIC) is a modern secure transport protocol designed to enhance the performance of TCP for internet communications. QUIC has been recently deployed to enhance perceived QoE [50], [51], [52]. It offers several advantages, including reduced latency [53], improved congestion control, and faster connection establishment.

2) NON-HAS-BASED STREAMING

Early but significant efforts in low-latency streaming technology led to the Real-Time Messaging Protocol (RTMP). RTMP is a TCP-based protocol that provides a stable connection and offers low-latency streaming capabilities, helping to mitigate buffering issues. However, RTMP suffers from poor quality, scalability, and passing through firewalls due to inadequate video transmission security. It is not natively supported by most end-user devices. RTMP started with the plugin of Adobe Flash Player. For two decades, RTMP was the defacto standard for video transmission over the Internet, but Adobe stopped supporting Flash Player at the end of 2020. Additionally, RTMP is vulnerable to bandwidth issues and fluctuations.

Real-Time Streaming Protocol (RTSP) is an older low-latency streaming technology. Its main advantage is the ability to provide segmented streaming, allowing clients to start playing videos without completing the download. RTSP achieves high throughput by using a TCP connection eliminating the need for local caching mechanisms. RTSP requires a connection to a dedicated server to stream video, which makes it difficult to scale. In addition, it does not support content encryption. WebRTC is another ongoing streaming technology that provides Real-Time Communication (RTC) by leveraging peer-to-peer connections between clients, enabling near-simultaneous data exchange. WebRTC provides low-latency streaming without requiring additional plugging. WebCRT adapts to network conditions, but it is not specifically designed with scalability in mind.

3) HAS-BASED STREAMING

HTTP adaptive streaming is a technology that takes into account network conditions intending to provide high-quality streaming. Easy scalability and compatibility with firewalls, as well as the sufficiency of the client to request each segment independently, represent the key reasons for using HAS technology in OTT media delivery. In HAS-based streaming, multiple copies of the same video are replicated at different quality levels for distribution. Players can then adaptively switch to the most suitable quality, optimizing the playback experience. In the following section, we will delve deeper into the intricacies of HAS video streaming.

IV. HTTP ADAPTIVE STREAMING (HAS)

A. ADAPTIVE VIDEO STREAMING TECHNOLOGIES

Adaptive streaming is a technology designed to provide high-quality streaming while considering network conditions. At a higher level of abstraction, in the adaptive streaming multiple copies of the same video are replicated with different qualities for distribution to the players, which can adaptively turn to a suitable quality. Three main components of adaptive streaming technology include: i) server, which provides live or on-demand streaming and has a responsibility to offer features like watermarking and DRM, ii) content distribution networks, and iii) player on the client side to playback the video.

Fig. 3 shows the higher-level abstraction of adaptive streaming. As seen, all streaming technologies (e.g., HLS, MSS, and DASH) follow the same principle. The basic idea behind adaptive streaming is generating multiple copies of the same content with different bitrates or representations. Each representation is divided into small equal-sized (or variable size) partitions (e.g., two, four, or six seconds of encoded video data) known as segments or chunks. All segments are stored on an origin media server and can be easily downloaded via a GET request. The origin server may also store an index file called a *'Manifest'*. The manifest file contains metadata about the video (e.g., the number of video and audio representations, subtitles, etc.) and details the media structure. It describes all relevant information about the representations and correlates chunks from different representations, as well as their corresponding Uniform Resource Locations (URLs).The origin media servers use media packager for packaging incoming requests [54], [55] taking into account multiple bitrates and media formats (e.g., HLS, MSS, and DASH), while automatically scaling outputs in response to audience demand. In the context of video delivery, HAS leverages CDN technology and the enormous benefits that CDN offers for QoE management in the current and future network landscape.

The packager segments the streams into small-size files and either pushes them to a CDN (push model), or stores them in the origin server so that CDN can pull from the origin server (pull model). Streaming video is delivered over the Internet using HTTP, allowing users to access the same content from different video devices (clients). Video playback begins with the request and download of the manifest file (e.g., .mpd, .m3u8). Subsequently, the player requests segment (s, b); segment *'s'* with bitrate *'b'*; based on manifest information, available network bandwidth, and the client's resources such as CPU utilization and buffer fullness. It is worth emphasizing that for each individual segment, adapting to a specific representation is done on the player side. The advantages of adaptive streaming technologies provide players with greater flexibility to enhance the overall QoE. Other concerns regarding the significant impact of adaptive technology on video streaming include network technology and implementation.

B. ADAPTIVE VIDEO STREAMING TERMINOLOGY

To better understand the concept of adaptive streaming, let's clarify some terminology.

- *Video Model*: In adaptive streaming, a video file is divided into N segments, which can be of equal or unequal size, indexed as $1, 2, \ldots, N$. Each segment represents p seconds of video, encoded in a different bitrate, refereed as representation. Higher bitrate representations offer better display quality but consume more bandwidth and have larger file sizes. A media file containing segments encoded at different bitrates and resolutions is housed on an HTTP media server.
- *Video Coding*: Video coding is an algorithm that compresses a video signal, most commonly based on Discrete Cosine Transform (DCT) coding and motion compensation. This process significantly reduces the size of the raw video recorded by the camera. Video codec refers to specific software, firmware, or hardware implementations capable of compressing (or decompressing) a specific video coding format. Most encoders support Variable Frame Rate (VFR) and High Dynamic Range (HDR) video formats.
- *Video Player*: A player is an application running on video devices (client) to playback video files. It downloads the manifest file and subsequent video segments from a CDN edge (cache hit) or the media server (cache miss) and plays them once the download is complete. In each adaption, a player only requests one segment from a media server. Also, the player has a limited buffer size (e.g., 40 seconds in length) to store downloaded video segments before playback. In this study, the terms 'client' and 'player' are used interchangeably.
- *Switch-over*: In adaptive streaming, which relies on bitrate adaption, the term *Switch-over* refers to transition among different representations, resulting in quality oscillation. The QoE is influenced by both video quality, measured by bitrate, and playback characteristics, such as rebuffering and switch-over. Therefore, adaptive streaming requires a tread-off between enhancing display quality and minimizing video re-buffering as well managing the frequency of switch-over.
- *Adaption Schemes*: Common adaption algorithms can be classified into three main predicted schema: client-driven adaptation, server-driven adaptation, and network-assisted adaptation. In recent years predicted adaptation schemas have been replaced with machine learning approaches. We will dive into each adaptation algorithm in the next section.

C. ADAPTIVE STREAMING PACKAGING FORMATS

Today, most content providers are still making four silos of content; HLS, MSS, DASH, and CMAF. While DASH

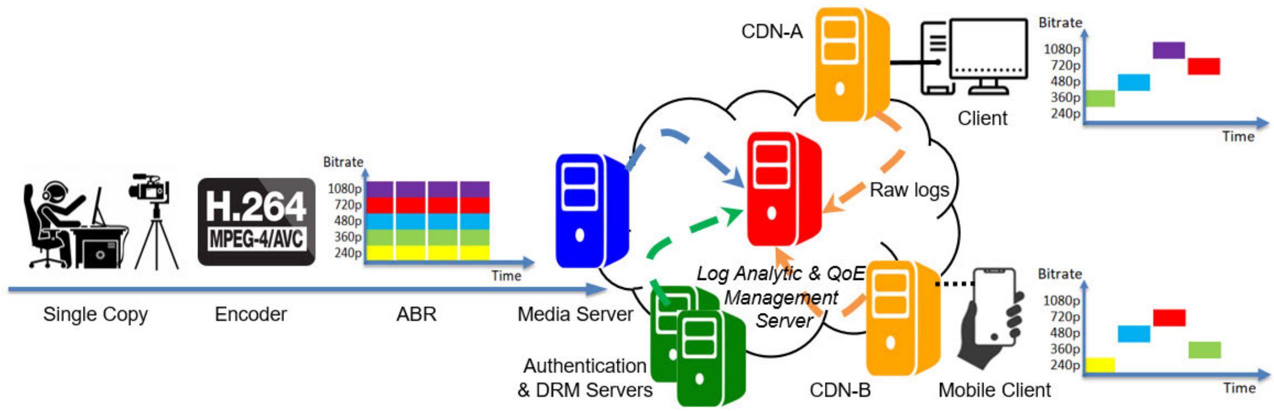


FIGURE 3. Abstract overview of the HAS technology and video delivery components, includes media server, content delivery networks, and client or media player. At the beginning of video streaming, clients authenticated by customer service management system, and short token assign for legally clients. A client without token (fraud user) can not access the video content.

primarily benefits from ISO Base Media File Format (ISOBMFF), the smooth streaming file format heavily relies on fragmented MP4 (fMP4) files, and HLS uses transport stream (.ts) files. As a result, content providers must encode and store multiple copies of video and audio files for the same content. Also, ingesting different codec formats from the same origin server negatively impact both the performance of the origin server and cache efficiency at the edge. Storing a single format within the CDN edge cache not only reduces cache fullness but also enhances the hit ratio. While HLS and DASH, are widely used by OTT providers, MSS is no longer supported by newer clients. However, some OTT providers, particularly those serving legacy clients like "set-top boxes" support the MSS format.

1) HTTP LIVE STREAMING

Adaptive streaming technologies are the ways to deliver multiple bitrates over HTTP, allowing players on the client side to automatically detect internet throughput and adjust to a suitable bitrate (or quality) in response. Unlike RTMP, which can struggle to pass through firewalls, HLS is CDN-scalable and firewall-friendly. Additionally, its ability for clients to request each segment independently are key reasons why it is favored for OTT stream delivery. This simplifies integration with existing infrastructures. Hence, HTTP streaming has become a more popular solution in commercial deployments [56]. A key benefit of the HLS protocol is its scalability and compatibility with a wide range of devices and firewalls. However, one of the trade-offs with HLS is its lag time (latency), which can typically range from 15 to 30 seconds for live streams.

2) DYNAMIC ADAPTIVE STREAMING OVER HTTP

MPEG-DASH is the only standard for multimedia streaming on the Internet. Fig. 4 shows a conceptual overview of DASH media streaming, where clients connect to the CDN-edge and start video streaming by requesting and downloading the manifest file. Players start streaming by requesting a Media Presentation Description (MPD) file and measuring download time. By parsing the received MPD file, players learn about the characteristics of the media such as URL location, DRM information, and available representations for video and audio files. Based on the achieved information, the client requests a suitable alternative of the first segment with triggering *HTTP GET* request. clients then continues fetching subsequent segments using an adaptation algorithm. Adaptation algorithms switch between different video representations, taking into account metadata information in the manifest file, current network throughput, and buffer fullness.

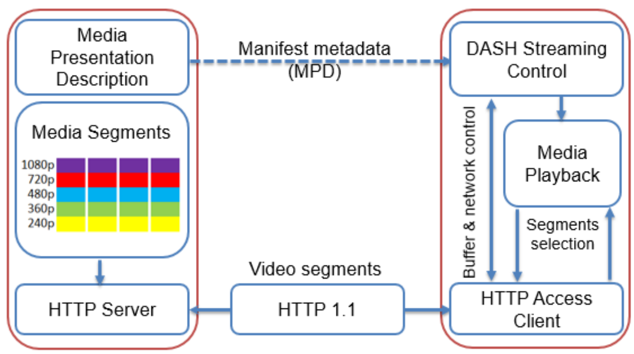


FIGURE 4. Abstract overview of MPEG-DASH player adaptation schema. The player begins streaming by downloading the MPD file and then continues by requesting the first segment. This process takes into account the manifest metadata and the download speed.

It has been proven that DASH client-pull technology is more flexible, firewall-friendly, and CDN-scalable than server-push technologies. However, the distributed and client-driven nature of DASH also causes some challenges. For instance, service providers may not be able to guarantee premium quality of service with DASH [57]. Because, the decentralized nature of the client's behavior, eliminates control over it. To overcome these challenges, MPEG introduced a new architecture referred to as Server and Network-assisted DASH (SAND) [58].

3) COMMON MEDIA APPLICATION FORMAT

Intending to provide a single format for distributors and playback that supports all types of client devices, Apple and Microsoft introduced the Common Media Application Format (CMAF) [59]. CMAF reduces encoding costs by eliminating the need for multiple file format and is compatible with HLS playlist (.m3u8) and a DASH manifest (.mpd). It supports HEVC/H.265, AVC/H.264, and Advanced Audio Coding (AAC). One of the main advancements of CMAF is its low latency. In live streaming, glass-to-glass delay is critical, specifically when content is transmitted via a satellite connection. In terms of security, it is possible to create an encrypted file conforming to different DRM standards simultaneously (Google Widevine, Microsoft PlayReady, and Apple Fairplay) with CMAF support.

D. BITRATE ADAPTATION SCHEMES AND QoE OPTIMIZATION

As previously mentioned, a glass-to-glass monitoring video delivery system enhances QoE optimization across various endpoints. Depending on the decision point for adaptation and requesting the next segment, bitrate adaptation schemes can be classified into three main categories, which we will discuss in detail.

1) CLIENT-DRIVEN ADAPTATION

Client-driven adaptation algorithms improve client QoE by downloading the highest possible representation while avoiding buffer starvation and oscillation among video representations. As shown in Fig. 5, the key features of all bitrate adaptation algorithms rely on client-side estimation. which includes the following:

- *Connection*: Estimates current network transfer rate and bandwidth fluctuations while also tracking the average network bandwidth.
- *Buffer State*: Current buffer fullness and average buffer fullness also affect client adaptation.
- *History*: Maintains a history of the current and average received bitrate.

2) SERVER-DRIVEN ADAPTATION

Client-driven adaption has some drawbacks. While service providers may not necessarily have control over the client's behavior, they may not be able to guarantee the premium quality of service with client-based adaptation. In addition, in client-driven adaptation, clients' devices may not be powerful enough to run traffic shaping and machine learning algorithms. Server-driven adaptation algorithms aim to assign fair quality where multiple clients compete for the available bandwidth. These algorithms enhance clients adaptation by sharing network information with clients. However, although servers store and manage client information individually, they

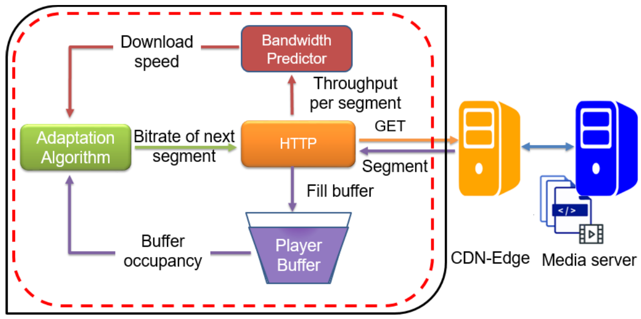


FIGURE 5. Bitrate adaptation schema in HAS video streaming. The adaptation algorithm takes into account the manifest file, download speed, and buffer occupancy.

come with significant drawbacks, including high overhead and complexity associated with server-based algorithms.

3) NETWORK-ASSISTED ADAPTATION

Client-driven adaptation, which relies on network throughput estimation, may be sub-optimal in large networks. It is not the preferred solution due to the inability to ensure optimal delivery under various large-scale system conditions. Leveraging centralized monitoring and management mechanisms can enhance clients to achieve better QoE in a shared competitive bandwidth. Pure network-assisted adaptation algorithms are monitoring systems that perform adaptation based on network data and statistics, independent of clients' data. Network middle-boxes, such as proxy servers, can assist clients during bitrate adaptation by providing network statistics and information like available bandwidth or suggestions for suitable bitrate.

E. SUMMARY

The main objective of this study is a survey recent machine learning and network-assisted adaptation schemas. However, at the end of this section we have a short list of client-side, server-side, and legacy network-assisted adaptation algorithms as shown in Table 5.

V. SERVER SIDE OPTIMIZATION AND QoE MANAGEMENT

A crucial question in the design of any video streaming system is; where should the decision point be placed to optimize application quality while considering limited network resource? As seen, most video streaming systems rely on a source-centric approach, where the content source determines the optimal compression and streaming methods. The media server encodes a video into multiple representations and distributes it through the delivery network, while the client display it in a specific format. However, the client does not typically provide any user-perceived QoE feedback to the media server.

Some research efforts focus on machine-centric video streaming [92], there camera (as a source) determines which frames and pixels to stream. The study advocates that the

TABLE 5. Classification of some state-of-art HAS adaptation schema.

Ref.	Method	Adaptation/ Description
[60]	SFT	Client-side /Available bandwidth adaptation
[61]	ESFT	
[62]	PANDA	
[63]	PiStream	
[64]	DASH2M	
[65]	LOLYPOP	
[66]	GeoStream	
[67]	BBA	Client-side/Buffer fullness adaptation
[68]	BIED	
[69]	BOLA	
[70]	QUETRA	
[71]	PID	Client-side/Hybrid adaptation
[24]	GTA	
[72]	MPC	
[73]	SARA	
[74]	SQUDA	
[75]	ABMA+	
[72]	FastMPC	
[76]	ELASTIC	
[77]	FESTIVE	
[78]	Pensive	Client-side/MPD -based adaptation
[79]	mDASH	
[80]	D-DASH	
[81]	LTS	Server-side/Global view adaptation
[82]	QAC	
[83]	QCSS	
[57]	SAND	
[84]	MS-Stream	
[85]	AVIS	Network-assisted/Legacy network adaptation
[86]	NOVA	
[87]	QARC	
[88]	FINEAS	
[89]	QDASH	
[90]	BUFFEST	
[91]	MP-DASH	

video streaming protocol should be driven by real-time feedback from the server-side Deep Neural Networks (DNNs). Despite the significant achievements in network resource utilization and quality improvement on the client side (camera) in an emulated environment, these approaches are largely impractical for real-world scenarios where hundreds of thousands of users simultaneously view the same video but provide different feedback. Furthermore, waiting for client feedback result in more latency which is critical for live streaming.

A. MULTI-PASS AND PER-TITLE ENCODING

Video encoding is the process of compressing video data using different standards or formats. After block partitioning, which divides the video into smaller blocks for more efficient processing, the following processes applied to the raw video. i) Transformation converts data from the spatial domain to the frequency domain and retains only the most significant frequency components (remove less important data). ii) Quantization reduces the precision of the transformed coefficients to minimize the bitrate. iii) Motion estimation encode one frame in terms of another by analysing motion. iv) Entropy encoding is a lossless data compression scheme that replaces data elements with coded representations.

Moving toward high resolution and frame rate brought the need for a more robust video codec that can offer fast and more efficient compressing. Optimizing video coding enhances end-users' perceived QoE and minimizes costs for OTT service providers. The primary of coding efficiency is the ability to encode video at the lowest possible bitrate while maintaining a certain level of video quality. On the server side, encoding a video with a higher bitrate, more representation, and a high-efficiency compression format results in time complexity and energy consumption. On the client side, downloading and decoding video before playing requires more time and processing power, which can impact QoE. This is particularly evident in wireless networks, where clients have limited processing power. Fig. 6, illustrates common codec distribution in the real world. HEVC/H.265 offers efficient compression compared with AVC/H.264, but significantly increased time complexity.

HEVC/H.265 utilizes a Coding Tree Unit (CTU) structure for block partitioning (allowing for blocks of up to 64×64 pixels). Each CTU can be recursively sub-partitioned into smaller square Coding Units (CUs) up to three times, with the smallest block size being 8×8 pixels (Fig. 7). The encoder traverses all Prediction Unit (PU) modes - both intra-prediction and inter-prediction- and determines the optimal one for each CU based on the rate-distortion function. This block partitioning scheme enables HEVC/H.265 to achieve more precise motion compensation, but at the cost of increased time complexity. In multi-rate encoding approaches, when a CTU is encoded, its information can be reused in other representations to minimize redundant search efforts.

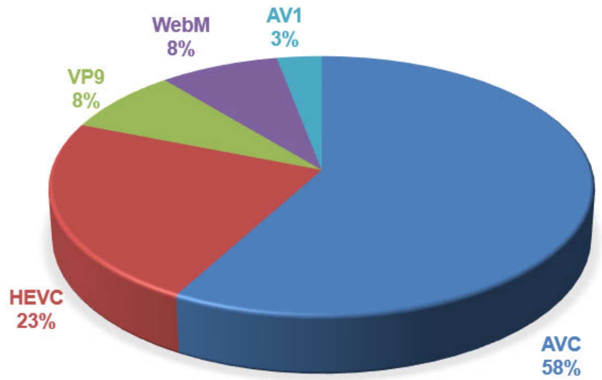


FIGURE 6. In 2022, Telestream/Encoding.com reported codec usage showing AV1 achieved a 30% improvement in compression over HEVC/H.265, which itself offers a 50% improvement compared to AVC/H.264. However, many legacy clients do not support AV1 or HEVC/H.265. In contrast, AVC/H.264 strikes a balance between efficient compression and high-quality playback, making it widely useful for video streaming, with support from nearly all clients, including legacy ones.

The efficiency of using a fixed, one-size-fits-all solution is reduced when dealing with diverse content complexities, varying network conditions, and different user device resolutions, potentially resulting in sub-optimal QoE. *LALISA* [93] is an efficient framework designed for dynamic bitrate ladder

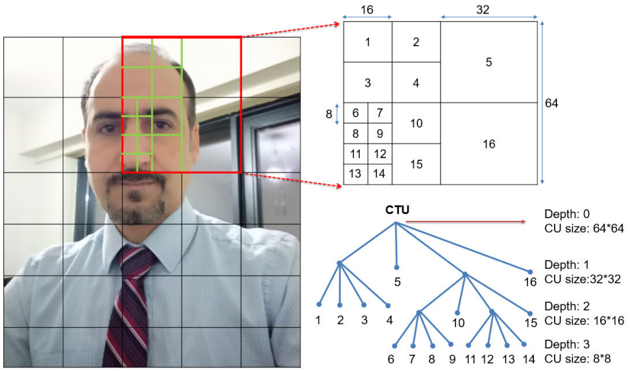


FIGURE 7. The red square denotes one CTU with a size of 64×64, and the digits within CTU represent the coding order of each CU. As illustrated, the red CTU is divided into 16 CUs, with the largest size of 32×32 and the smallest size of 8×8.

optimization within live HAS setups. It can be deployed along the path between client and the origin server. LALISA adapts the bitrate ladder of a live video session in real-time, resulting in enhanced perceived QoE and saving encoding, storage, and bandwidth costs. *ARTEMIS* [94] represents a practical and scalable alternative that dynamically adjusts the bitrate ladder based on factors such as content complexity, network conditions, and client statistics. It seamlessly integrates into the end-to-end streaming pipeline, operating transparently for both video encoders and clients. When significant fluctuations in available network bandwidth increase the likelihood of stall events, ARTEMIS prioritizes lower bitrates in constructing the ladder. This approach aims to maintain QoE as close as possible to the maximum achievable, even under adverse network conditions. It also reduces encoding computation, decreases storage and bandwidth costs, by scaling down the bitrate ladder.

A typical predefined bitrate ladder is employed for live streaming without considering optimization. However, a content-aware approach, known as per-title encoding, has been introduced to address video content characteristics and network conditions. This method can improve the QoE while reducing the bitrate. For example, in the two-pass encoding schema, the input video content is analyzed in the first-pass to inform the second pass, allowing for better encoding decisions and improve overall compression efficiency. Although this method is effective for VoD applications, it is not applicable for live streaming where a single-pass encoding schema is mainly used to minimize time complexity. On the other hand, a predefined bitrate ladder enhances simplicity and efficiency by eliminating the need for additional run-time to determine the optimal pairs (bitrate-resolution) during the live streaming session.

State-of-the-art methods prioritize encoding the highest quality representation first and reuse information from the encoding process to accelerate the coding of the remaining representations. However, this method can become a bottleneck in parallel encoding scenarios, as encoding the highest quality representation typically requires higher time complexity compared to lower quality representations. As a result, the overall time complexity is often limited by the time required to encode the highest quality representation. Research in [95] argues that encoding a middle-quality representation as a reference can substantially reduce the maximum encoding complexity, making it an efficient approach for parallel encoding of multiple representations. In *FaME-ML* [96], fast multi-rate encoding for ABR streaming using Convolutional Neural Network (CNN) is introduced. The emphasis is on reducing the highest time complexities rather than uniformly decreasing complexities across all representations, aiming to reduce overall time complexity.

In [97], a CNN is leveraged to accelerate multi-rate and multi-resolution encoding for ABR streaming. In multi-rate encoding, the lowest bitrate representation is selected as the reference for the neural network to optimize encoding efficiency. In multi-resolution encoding, the highest bitrate of the lowest resolution representation is used as the reference. In HEVC/H.265, for dependent representations, pixel values from the target resolution and encoding information from the reference representation are employed to predict CTU split decisions. The proposed method in multi-rate encoding and multi-resolution encoding can reduce the overall encoding time compared to the HEVC/H.265 reference software.

B. SECURITY IN VIDEO STREAMING AND QoE MANAGEMENT

The presented video security algorithms exhibit several limitations and drawbacks, including computational complexity, lack of robustness, significant delays, excessive consumption of memory resources, absence of comprehensive security, and insufficient privacy protection due to their simplistic design [98]. Security issue in media streaming can be classified into five principles; i) content confidentiality, ii) content integrity, iii) content availability, iv) user authentication, v) and DRM. In this section we will focus on user authentication and DRM. Generally, the presented video security algorithms focus solely on achieving the confidentiality of transmitted videos without addressing copyright protection, or vice versa. While DRM ensures that only authorized users have access to the appropriate content, most of today's watermarking technologies enable unique content requests for each client. This facilitates tracking of content distribution from anonymous sources (pirate) [99]. Watermarking serves as a method to protect intellectual property and provides proof of ownership in cases of piracy. In order to support traceability, a watermarking scheme must meet certain requirements. First, it is essential to ensure that embedding the watermark does not affect video quality and remains undetectable to human observers [100]. Second, the watermarking scheme has to be robust enough against geometric, compression and camcorder recordings attacks.

Finally, the application of watermark should not negatively affects video packaging, delivery, and playback time.

According to the architecture of the video delivery system, two types of watermarking solutions are common: i) Client-side watermarking: The user's device has an embedded agent integrated with the application that performs the watermarking, allowing it to create a unique request for content. ii) Server-side watermarking: The server is responsible for the watermark operation. The server side (head-end) watermark (a type "A" and "B") is transparent to clients, meaning that no changes are required on the client side. Unlike client-side watermarking, which is typically embedded in the video with an image overlay created by either the client or server, server-side watermarking is embedded in the video stream before it reaches the client device. It is important emphasizing that, the end-user devices have limited processing resources to generate an image, so most client side watermarks are content-unaware (the ability to determine the information contained in a video frame) [101]. Furthermore, compared to the client-side (multi point configuration) watermarking that requires DRM triggering, client hardening, and separate integration for current and newly introduced client types, single point head-end watermark solution reduces the risk and complexity of continuous integration.

When the client accesses the watermarked content, interactions occur between the various components involved in the watermarking process. Fig. 8 provides a simple overview of the functional architecture of HAS technology and its forensic watermark implementation. Forensic watermarking makes enables the detection of pirates who illegally redistribute copyrighted videos. For adaptive streaming, these methods are most effective when combined with "A/B" watermarking, in which two watermarked versions are created for each video segment and subsequently combined to produce a large number of uniquely watermarked videos [101]. Before starting to playback the video, the client must first be authenticated by the Content Management System (CMS). The CMS authenticates the user and assigns a short token to authorized legitimate users for connecting to the Edge-CDN. To initiate streaming of video content, clients request DASH 'manifest' or HLS 'playlist' files from the CDN. These files, which contain information about "A" and "B" variants, are provided to all clients. However, each client receives a unique sequence of "A" and "B" variants. Therefore, before requesting segments from the CDN, each client requires an authorization token and a watermark token. The watermark token includes a watermark pattern and a set of decisions regarding "A" or "B" variants. At the edge, the CDN manages client requests and determines which variant ("A" or "B") should be assigned. With each request, the CDN first checks the edge cache; if the content is not cached, the request is then forwarded to the origin server.

In [102], watermarks are embedded at specific locations, and a location map is used to help the encoder locate the watermark at the same position, as the location may change after compression. However, this method has security

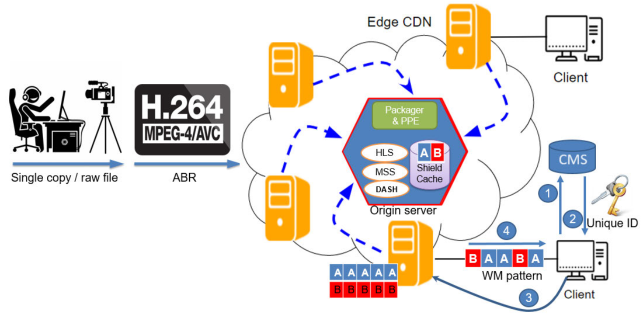


FIGURE 8. Integrated schema of ABR streaming with session based forensic "A/B" watermarking. A different combination of "A" and "B" makes unique request for each user [101].

drawbacks, as these locations are vulnerable to detection by attackers. To solve this problem, the method proposed in [103] applies the texture and motion information of the video content to realize the optimal watermark embedding location adaptively. The candidate blocks in the optimal location offer better robustness against re-compression attacks compared to other regions. The need for original data during watermark extraction (context-aware algorithms) is usually simpler and more effective. However, in most applications, the original host video is not available for watermark extraction. Therefore, the watermark must be identified without relying on the original video content (context-insensitive algorithms). The authors in [100] demonstrate that by designing a hybrid selection algorithm using texture information and feature points, the method automatically selects stable blocks can avoid being destroyed during video encoding and complex attacks. As a result, high video quality and robustness against geometric, compression, and camcorder recording attacks are achieved.

Watermark either before encoding (pre-encoding) or after encoding (post-encoding), can lead to high computational complexity and poor real-time performance. The implementation-free, rate-distortion-preserving, and scalable "A/B" watermarking concept introduced in [104]. The advantage of this method is that it does not require to change the existing video encoder implementation. Instead, it only modify the target bitrate parameter to create different compression artifacts. These artifacts represent watermarks but are not noticeable due to high-quality encoding. As a result, the robustness is equal to or better than state-of-the-art methods with comparable embedding complexities. Furthermore, the rate-distortion performance is roughly equivalent to conventional compression without watermarks. The authors argue that the proposed method can be employed in practice without implementation overhead or compromising quality.

Authors in [101] implemented a head-end post-packager watermark solution on a real-world OTT platform and found that legacy devices, typically using the MSS format, struggled to meet watermark requirements. Furthermore, applying a watermark added extra processing load to the origin server because it involved preparing two different variants (''A'' and ''B''). Due to server storage limitations, network bandwidth bottlenecks during transmission, and terminal equipment performance limitations, the server may need re-compress the video until it meets the publishing rules requirements (e.g., format and size), before uploading it. Therefore, developing a video watermarking algorithm capable of resisting re-compression attacks has become a critical issue, especially when the quantization parameter is significantly increased.

Despite the efficiency of the HEVC/H.265 codec, it lacks reliable and practical features for supporting authentication and copyright applications. To address this gap, researchers have proposed various watermarking techniques in recent years. The high-fidelity HEVC/H.265 protection and verification framework suggested in [105] employs a discrete transform technique for inserting internal block-based watermarks into other blocks of the transmitted HEVC/H.265 frames. The results demonstrate the feasibility of watermark protection and integrity verification while achieving high robustness against multimedia attacks and channel noise. A semi-fragile watermarking algorithm based on chromatic residual Discrete Cosine Transform (DCT) of AVC/H.264 is proposed in [106]. The DCT coefficients are divided into ''DC'' and ''AC'' coefficients. The study found that 'chroma' blocks exhibit a more stable prediction mode compared to 'luminance' blocks. Therefore, to enhance video quality and mitigate bitrate increases, the watermark is embedded in the chromatic intermediate frequency quantized AC coefficients. The results demonstrate high sensitivity to malicious attacks in both the spatial and temporal domains.

While many existing video watermarking schemes require synchronization to extract watermarks from geometric attacks (e.g., rotated, scaled, and cropped videos), which can be time-consuming and impact accuracy, the method proposed in [107] allows for direct extraction without the need for synchronization, even when videos are subjected to rotation, scaling, or cropping. In this method, circular templates in the Discrete Fourier Transform (DFT) domain are transformed into spatial masks and added to video frames in the spatial domain. A local contrast-based perceptual model is applied to maintain the fidelity of watermarked video. Early effort to embedding and retrieving information in HEVC/H.265 as well as the aforementioned studies [106], [107] rely on modifications of standard, making it difficult to deploy them into commercial systems. The method proposed in [108] address this challenge without compromising quality. To this end, the method relies on changing the luminance of specific blocks instead of making alterations within the HEVC/H.265 codec itself. However, the proposed method deals more with steganography rather than watermarking, as it embeds a watermark as a sequence of unique symbols for each user.

In recent years, a variety of study focused on leveraging the advantages of blockchain technology in watermarking using ''zero'' watermarking, which involves no alterations to the original video. However most of those works primarily consider image watermarking. A blockchain-based zero-watermarking algorithm combining Non-Subsampled Contourlet Transform (NSCT) with Singular Value Decomposition (SVD) is introduced for video protection in [109]. This approach addresses rights traceability and enables tamper-proof storage by integrating blockchain technology with zero-watermarking. In [110], watermarks are embedded based on previous frames, forming a chain-like structure that is crucial for determining whether a frame has been tampered with or not. If someone attempts to modify a frame, the hash of the previous frame will no longer match the extracted watermark. Therefore, to alter one or more frames in the video, the attacker would need to modify all frames preceding the altered frame(s). The algorithm cannot detect audio tampering that occurs alongside the video.

Machine learning approaches for watermarking are introduced in [111] and [112]. These studies present a method for watermarking MPEG-4 videos using Extreme Learning Machine (ELM) to select frames, along with a combination of fuzzy logic and particle swarm optimization for embedding and extracting watermarks. As a results, the signed videos reflect good visual quality with high PSNR values and are robust against video processing, geometry, and noise attacks. Based on our experience with real-world OTT platforms, implementing any watermarking solution adds processing and delivery overhead to the video streaming system. Additionally, applying watermarks uniformly across all clients can be challenging due to varying technical requirements and capabilities.

Overall, watermarking is used to trace the source of content redistribution (piracy). It involves embedding symbols into content to make it difficult for pirates to detect or extract them, as attackers attempt to analyze and extract these embedded symbols. Three key factors must be considered when evaluating a video watermarking system: The impact of the watermark on the visual video or audio (perceptual quality), the strength of the watermark against unauthorized removal or alteration (security), and its detectability after various distortions or transformations (robustness). These factors are often in tension with each other, meaning that improving one factor can often lead to compromises in the others (e.g., enhancing robustness may reduces visual quality). A systematic literature review in [113] evaluates research conducted between 2014 and 2020, identifying challenges within HEVC/H.265 and suggesting potential avenues for improvement:

- A trade-off between complexity and security should be maintained.
- Considering hardware-based solutions to support real-time applications.
- Error detection and correction codes can further reduce the occurrence of errors when highly robust watermarks are required.

- All potential attacks that could affect the readability of watermarks should be considered to ensure they can strongly resist such threats when high robustness is necessary.

C. SUMMARY

Table 6 summarizes all the server-side optimization and QoE management papers discussed in this section.

VI. EMERGING APPLICATION AND ENERGY AWARE TECHNOLOGIES

Video streaming has become a major usage scenario for the Internet. The increasing popularity of mobile devices and new applications, such as 4K and 360-degree videos, require careful sharing of network resources among users to achieve the QoE and fairness objectives. This section discusses emerging applications and network trends that enhance fairness and QoE management, including ML, CDN, SDN, and NFV technologies.

A. EMERGING APPLICATION AND QoE MANAGEMENT

Recently, we have witnessed a rapid increase in the use of machine learning techniques and artificial intelligence in various domains. These emerging applications are increasingly being used to improve and automate processes related to monitoring, data analysis, predicting user behavior, and optimizing services. In the context of QoE management for video streaming, ML and AI play an important role in improving the overall user experience. They achieve this by dynamically adjusting parameters such as video bitrate, content optimization, network routing, and resource allocation based on real-time client behavior and network conditions.

So far, a significant number of studies have adopted ML approaches for QoE management. The limitations of state-of-the-art solutions for QoE management is highlighted by [114], which points out that using a fixed set of rules may not always guarantee optimal video quality and accurate buffer estimation, particularly during unpredictable bandwidth fluctuations. To handle these issues across a diversity of networks, the study proposes a Deep Neural Network (DNN) that selects the appropriate bitrates to maximize the overall user experienced quality. Comparison results with *Pensive* [78] as a state-of-the-art solution, indicate that the proposed machine learning method achieves superior performance.

State-of-the-art video adaptation algorithms heavily rely on network bandwidth estimation and often do not integrate video quality enhancement techniques or consider the user device resources. Consequently, this approach results in suboptimal QoE. Downloading segments at a higher bitrate usually provides better video quality but may result in buffering. On the other hand, lower bitrates can be downloaded faster, but with lower quality. In *SRAVS* [115], Reinforcement Learning (RL) model is applied to integrate the super image resolution technique with video streaming strategy. The RL model analyzes the playback statistics and the distinguishing features related to the client-side computing capabilities. The video super-resolution technique allows clients to download video segments with a lower bitrate and upscale them to high-quality resolution, reducing the system's reliance on dynamic bandwidth estimation. Although SRAVS can significantly improve the QoE for users compared to the state-of-the-art video streaming strategies, it is not compatible with legacy devices due to a lack of processing power.

The authors in [116], leveraged a reinforcement learning ABR algorithm to enhance QoE in DASH streaming. For this purpose, the ABR selection problem is formulated as a Markov Decision Process (MDP) problem. Accordingly, an RL-based solution is proposed to solve the MDP problem, in which the DASH clients act as the RL agent, interacting with the network dynamics as the environment. The proposed algorithm jointly considers video quality and buffer fullness and aims to dynamically select the optimal bitrate for each video segment, thereby increasing the overall video quality.

Leveraging combination features of RL and VNF is introduced in *Rldish* [117] to enhance QoE. In this study, VNF is deployed in a real HTTP cache server at the edges, and its performance is evaluated using streaming servers distributed globally. Compared with the state-of-the-art solutions, the proposed method is lightweight and fully transparent to both clients and the streaming server. *Flex-Steward* [118] performs fair bandwidth allocation among clients in the shared bottleneck link by training an adaptive bitrate delivery algorithm based on Neural Networks (NN) and RL at the network edge. Similar to *Rldish*, this method achieves superior performance compared to state-of-the-art solutions.

Most developed ML models have three main limitations: i) Existing models rely on features that are unique to a specific dataset, and thus cannot be generalized. Due to the complex interactions between these features, there is a need for a unified representation that is independent of datasets with diverse modalities. ii) Some models often lack the configuration capability to perform classification and regression tasks. iii) Many existing models depend on small-sized datasets, and the impact of limited data on the performance of QoE models has not been sufficiently addressed. *DeepQoE* [119] is an end-to-end framework that uses deep learning techniques to address these challenges. During the learning stage, neural networks assist in extracting generalized features. The learned representation then serves as input for classification or regression tasks.

Server-side reinforcement learning approach discussed in [120] aims to provide fair QoE among clients sharing a bottleneck bandwidth. In this model, the sever selects a bitrate for each DASH client by modifying the client's MPD file. While server-side adaptation algorithms require storing more information, which can negatively impact server performance, the study proposes a Recurrent Neural Network (RNN) to minimize the number of actions needed for the server to maintain the system in equilibrium.

TABLE 6. Summary of server-side QoE management schema.

Ref.	Description	QoE	Protection	Legacy	ML	Blockchain
[92]	Deep learning models offers information that helps better video encoded/streamed	✓	✗	✗	✓	✗
[93]	A lightweight, codec-agnostic bitrate ladder for optimizing HTTP live streaming	✓	✗	✓	✗	✗
[94]	Enhance QoE and encoding time by optimizing bitrate in relation to network bandwidth	✓	✗	✓	✗	✗
[95]	Reduce encoding time complexity by using a mid-quality representation as a reference	✓	✗	✓	✗	✗
[96]	A ML approach that uses low-bitrate coding information to encode higher bitrates	✓	✗	✗	✓	✗
[97]	A machine learning approach for fast multi-resolution and multi-rate encoding	✓	✗	✗	✓	✗
[100]	Implementing watermark that utilize adaptive region selection and channel referencing	✗	✓	✗	✗	✗
[101]	Implementing watermark in real live video streaming platform	✓	✓	✗	✗	✗
[102]	Robustness and quality of embedding watermarks in HEVC/H.265	✓	✓	✗	✗	✗
[103]	Implementing watermark against re-compression attacks utilizing texture & motion inf.	✓	✓	✗	✗	✗
[104]	Implementing forensic & scalable watermark for video streaming without rate-distortion	✗	✓	✓	✗	✗
[105]	Integrating watermarking and steganography to ensure confidence in HEVC frames	✗	✓	✓	✗	✗
[106]	A watermarking scheme using chromatic DCT for video protection and tamper detection	✗	✓	✓	✗	✗
[107]	A watermarking scheme that extracts watermarks in rotated/scaled videos without Sync.	✗	✓	✓	✗	✗
[108]	Hiding and retrieving information in high-resolution HEVC videos using steganography	✗	✓	✓	✗	✗
[109]	Blockchain zero-watermarking to improve robustness and address traceability challenges	✗	✓	✗	✗	✓
[110]	Watermarks are embedded from previous frames, forming a chain for detecting tampering	✗	✓	✗	✗	✓
[111]	Integrating fuzzy logic & particle swarm optimization for watermark embedding/extraction	✗	✓	✗	✓	✗
[112]	A robust deep learning-based watermarking scheme against modifications or distortions	✗	✓	✗	✓	✗

The limitation of heuristic approaches for serving client requests at the network edge in the presence of highly dynamic network conditions and diverse request patterns, which can lead to significant overhead and increased time complexity as addressed in [121]. The proposed learning-based client request management solution at the edge leverages advancements in the Deep Reinforcement Learning (DRL) to handle requests from concurrent users accessing various HLS channels efficiently. Additionally, a joint optimization strategy aimed at achieving high QoE in live events without latency or buffering is discussed in [122]. In this study, a DRL framework is employed to maximize QoE for live streaming without relying on assumptions about the environment or fixed rule-based heuristics.

Although DRL approaches have been successfully applied to network optimization problems by modeling them as Markov decision processes [123], existing RL algorithms involving multiple agents struggle to address nonlinear objective functions in the rewards of different agents. To address this issue, *MAPG-finite* [124] introduces an optimization model for nonlinear objective functions that represent cumulative rewards of multiple agents within a finite time horizon. The primarily objective of this study is to enhance QoE and fairness among various users in a video streaming network.

Researchers at *Stanford* and *Tsinghua* University, who monitored streaming video for over 60K users, observed that in real-world conditions, sophisticated or machine-learned control algorithms often struggle to outperform a "simple" algorithm (e.g., buffer-based), despite demonstrating good performance in network emulators or simulators. According to statistical analysis, the heavy-tailed nature of network and user behavior, coupled with the challenges of emulating diverse Internet paths during training, presents significant obstacles for learned algorithms in this context. To address this gap, they introduced *'Fugu'* algorithm [125], which is based on Model Predictive Control (MPC) [72]. This algorithm utilizes supervised learning in situ, using data from the real deployment environment, to train a probabilistic predictor of upcoming chunk transmission times. The *Fugu* achieves superior performance compared to state-of-the-art algorithms by leveraging data from its deployment and restricting the scope of machine learning to make predictions that can be quickly validated.

Data-driven schema (e.g., *MPC [72], Pensive [78], Fugu [125]*) can facilitate flexible QoE objectives, but they may suffer scalability issue. During the offline phase, these methods typically learn a representation of the state decision map (whether as a table or neural network) optimized for a predetermined QoE objective. In the online phase, decisions can be quickly made based on this learned representation. However, each representation typically supports one objective unless reconfigured, making it impractical to handle a large number of QoE diversity simultaneously. To address this limitation, the authors in [126] introduce *'Ruyi'*, a supervised RL approach. Ruyi's ABR algorithm is designed to directly predict the influence on metrics after taking different actions. With these predicted metrics, Ruyi adapts the bitrate that maximize user-specific QoE once the preference is given. As a result, Ruyi is scalable for different user preferences without the need to re-training the learned models for each user.

As stated, existing ABR algorithms face challenges in handling QoE diversity in two folds: i) Simple ABR algorithms often perform straightforward bitrate selection without effectively addressing QoE diversity. ii) Predictive models and ML-based approaches (e.g., *MPC* [72], and *Pensive* [78]) typically assume fixed QoE function. To address this issue, the authors in [127] introduced *Elephanta*, an online flexible ABR algorithm designed for edge users. Elephanta incorporates; i) A user QoE perception interface to better understand and adapt to individual QoE requirements. ii) An adaption algorithm with flexible parameters. To minimize the overhead of online update, video streaming is modeled as a renewal system, and QoE functions are formulated into flexible formats.

Recent studies have focused on understanding the characteristics of user behavior and how these factors influence QoE. Enhancing perceived QoE by monitoring end-user streaming behavior and fair bandwidth allocation is discussed in [128]. This study introduces a monitoring mechanism that traces users' download and display behavior, applying restrictions to users who download more segments without playing them. Additionally, the study presented in [129] research on social context factors and user engagement characteristics and investigate their correlation with QoE. Accordingly, build a metric that estimates the end-to-end QoE for a specific aspect of user actions. Then, applying ML models to predict QoE; finally, validate this approach using metrics for statistical evaluation of quality prediction models.

In [130], three different ML approaches (Neural Networks (NN), Long Short-term Memory Networks (LSTM), and Random Forest (RF)) are developed and compared to estimate playback behavior based on uplink request information in the given data set. This approach predict QoE parameters, including startup delay, rebuffering, video quality, and quality oscillation. While the NN approach is lightweight, it tends to exhibit a larger variance in estimation errors compared to the RF approach. In some instances, however, NN performance can be comparable to RF. On the other hand, the LSTM-based approach, demonstrates identical performance for the investigated metrics but introduces higher complexity and requires more computational resources. The RF approach performs well across all QoE metrics when tested on a known dataset, showing only slightly reduced performance on an unknown dataset.

A new approach to improving the stable quality of video streaming in the DASH using DRL is discussed in [131]. To enhance QoE, the proposed model dynamically adjusts the quality distance factor between consecutive video segments. In this model, the client-side adaptation algorithm determines bitrate decisions based on current network bandwidth, buffer states, and the quality of the previous video segment. Additionally, the research presented in [132] focuses on metadata learning strategies in DASH. The proposed framework includes data collection and preprocessing, meta-model architecture, meta-training, meta-testing, fine-tuning, and continuous improvement. Using prior knowledge and experiences, meta-learning enables DASH streaming to quickly and effectively adapt to changing network conditions and user preferences, ultimately improving the end-user's QoE.

Most ABR algorithms dynamically select the bitrate for each segment based on perceived network capacity and buffer occupancy. Unfortunately, these algorithms usually try to maximize the average bitrate rather than perceived quality, which leads to QoE degradation. To overcome this obstacle, the dynamic chunk quality-aware adaptive bitrate algorithm introduced in *DAVS* [133] selects a higher quality for dynamic chunks without significantly degrading the quality of static chunks. It also, takes into account the user's observation preference to adapt DAVS to variation in QoE.

Reinforcement learning algorithms optimize their control policies based on the actual performance of past choices, allowing them to discover policies that outperform algorithms relying on static heuristics. As mentioned earlier, some RL methods can learn the amount of buffer occupancy to reduce the risk of buffering based on the network throughput dynamism. Several previous studies have shown promising results for RL methods in controlled experiments with a fixed set of videos and network traces. However, it remains unknown how well RL-based methods are robust and perform well in large-scale real-world settings. To bridge this gap, *ABRL* [134], an RL-based ABR module, is introduced into Facebook's production web-based video platform. Regarding outperform achievement, this study only considers web clients, with no experimental results available for other client types. Furthermore, developing a similar learning framework for mobile and cellular networks, which often exhibit high dynamism, could potentially yield even greater ABR improvements.

The main challenge in evaluating HDR video quality lies in the expanded dynamic range and color gamut, which makes is default to create a single accurate quality metric. Video Multimethod Assessment Fusion (VMAF) is designed to correlate closely with human perception of video quality and is widely used in video encoding to optimize video compression and delivery. It utilizes an objective video quality assessment metric that integrates several individual metrics, including Detail Loss Metric (DLM), Visual Information Fidelity (VIF), and Mean Co-Located Pixel Difference (MCPD), to generate a single quality score. The supervised machine learning method proposed in [135] predicts VMAF scores for HDR. To train the ML algorithm, the study collected a set of HDR datasets and their corresponding Standard Dynamic Range (SDR) versions, along with intention video quality scores, especially VMAF values.

Most ABR schemes employ either manually tuned heuristics or learning-based methods. Heuristics are straightforward to implement but may not consistently deliver optimal performance. In contract, learning-based methods typically perform well but can be challenging to deploy on resource-constrained devices. To leverage the strengths of both approaches, *Ahaggar* [136] introduces a server-side learning-based scheme that offers quality-aware bitrate guidance to streaming clients using their own heuristics. The proposed schema utilizes the new Common Media

Client/Server Data (CMCD/SD) protocols to facilitate the exchange of essential metadata between servers and clients. While Ahaggar represents a significant advancement, the study in [137] focuses on three specific areas: i) Evaluating Ahaggar's performance in heterogeneous environments that include both Ahaggar-enabled and non-Ahaggar clients across diverse network conditions and device resolutions. ii) Quantifying the impact of device resolution on QoE when using Ahaggar. iii) Additionally, analyze the design choices behind Ahaggar. In conclusion, Ahaggar enhances user experience by achieving superior performance with reduced bandwidth consumption across diverse network conditions.

B. ENERGY-AWARE ADAPTATION AND QOE MANAGEMENT

Multimedia services consume a large portion of the Internet data traffic, highlighting the need for techniques to minimize energy consumption during both streaming and playback processes. Using perceptual quality metrics, such as VMAF, helps to understand the relationship between video bitrate and perceived quality. The trade-off between bitrate and quality is essential, as reducing the bitrate can lead to decreased power consumption in end devices. The ABR algorithms can result in substantial variations in power consumption, which is particularly crucial for mobile devices. In this regard, many studies have investigated energy-aware ABR streaming.

Multi-Codec Bitrate ladder Estimation (MCBE) [138] is an energy-efficient ABR streaming scheme. This approach involves two key components: i) multiple video codes are used concurrently for encoding a video with different bitrates, utilizing a lightweight codec such as AVC/H.264 for lower bitrate encoding and more efficient codecs like HEVC/H.265 or AV1 for higher bitrate encoding. ii) In practice, the VMAF scores of different representations are often very similar, leading to perceptual redundancy in the bitrate ladder. This redundancy results in wasted energy during encoding, storage, and transmission. To address this, a lightweight algorithm is proposed to remove redundant representations for each codec. Although this scheme results in efficient video playback, players can face challenges when smoothly switching between video codecs.

For mobile clients, energy consumption can be optimized from the perspective of video download. Due to vibration while moving, QoE decreases even when video segments are downloaded at higher bitrates. Thus, downloading segments with optimal bitrate in such environments may significantly reduce energy consumption, without greatly degrading the QoE. The context-aware video streaming algorithm in [139] aims to conserve energy in smart devices during movement time, meanwhile saving video quality at an acceptable level.

GreenABR [140] is a deep reinforcement learning model employs a standard perceived quality metric, VMAF, along with real power measurements collected through a streaming application. It learns how to adapt to network dynamism, maximizing perceived QoE while minimizing the power consumption on mobile devices during video playback (Fig. 9). *EnDASH* [141] uses a DRL-based neural network algorithm to minimize energy consumption in cellular networks. The proposed algorithm initially adjusts the playback buffer length based on the average predicted cellular network throughput and subsequently selects an optimal video chunk bitrate. This approach achieves a 30% improvement in energy consumption compared to the *Pensieve* [78] algorithm without compromising QoE.

The extended version of this study (*EnDASH-5G*) [142] utilizes a novel throughput prediction mechanism for millimeter-wave 5G (mmWave 5G) networks. This approach enhances existing throughput prediction models through a transfer learning-based methodology, leveraging publicly available 5G datasets. This adaptation allows EnDASH-5G to effectively predict throughput under the specific conditions and challenges posed by mmWave 5G technology. This synchronization ensures that data download attempts are aligned with the underlying network conditions. By accurately predicting throughput in 5G networks, the device can schedule data downloads during optimal network performance periods. This synchronization prevents unnecessary energy consumption during periods of poor connectivity or congestion, ultimately increases the device's battery life. Fig.10 shows a scaled spider chart across QoE components, where 10% corresponds to maximum value for each parameter. As seen, EnDASH-5G sacrifices average bitrate for stability, while *BOLA* [69] and *Pensieve* [78] focus on optimizing either minimal rebuffering or bitrate oscillation.

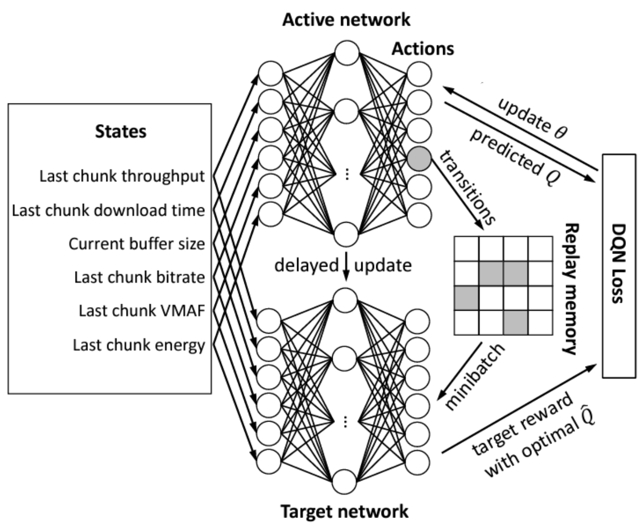


FIGURE 9. GreenABR agent uses a four-layer feed-forward neural network as the Deep Q-Network (DQN) to calculate the expected cumulative reward Q(s, a) for each state-action pair (s, a). The neural network weight parameter θ is learned from the feedback of the client environment [140].

DASH performance can be significantly affected by network conditions, so using multiple connections can help mitigate network uncertainties. Multipath techniques can simultaneously establish multiple subflows to

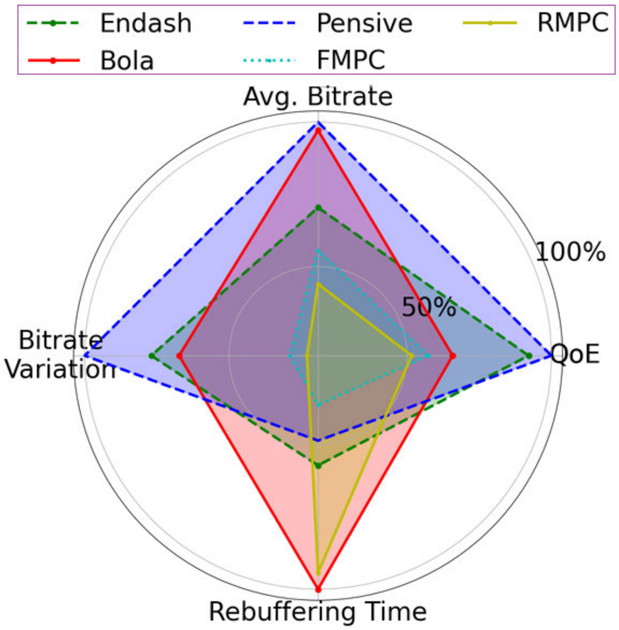


FIGURE 10. Due to tuning the buffer length according to the average predicted throughput, the average bitrate of EnDASH-5G is lower than BOLA and Pensieve. On the other hand, BOLA primarily uses playback buffer length to predict bitrates. This strategy often results in lower bitrate oscillation, but can lead to higher rebuffering times. Pensieve uses past chunk download rates to predict future bitrates. This captures the instantaneous throughput variation, reducing the rebuffering time but increasing the bitrate oscillation [142].

enhance throughput during the transmission and streaming high-quality video on (mobile) devices. However, current ABR video streaming systems have difficulty in multipath scheduling and balancing between QoE and QoS. To tackle this concerns, *PERM*, a neural adaptive video streaming with multipath transmission, [143] employs two actor modules and a critic module: the two actor modules respectively assign the path usage of each subflow and select bitrates for the next chunk of the video, while the critic module predicts the overall objectives. Results show that PERM outperforms both state-of-the-art multipath and single-path streaming systems, improving QoE and QoS metrics by 10%-15%.

In order to support the provision of high-quality 5G media services, researchers need to investigate Multipath TCP (MPTCP) inefficient data scheduling in heterogeneous subpaths, considering the energy consumption and its inconsistent behavior when using DASH application layer protocol. *QLE-DS* [144] is an MPTCP-based media service model used for multipath scheduling as a Q-learning process and employs a novel quantum clustering approach to discretize the high dimensional continuous Q-table. QLE-DS focuses on analyzing how path capacity and data retransmission affect energy consumption, enabling it to estimates the actual condition of each network path. This understanding helps QLE-DS to optimize flow completion time, which is critical for improving overall network performance.

The trade-off between energy consumption and QoE management is discussed in [145]. The proposed algorithm addresses the issue of network resource availability over time in a multi-user environment, aiming to save energy while ensuring accurate video quality in Log Term Evolution (LTE) and 5G networks. Content caching, encoding, and transmission for ABR streaming in cache-enabled cellular networks are studied in [146]. This study considers the dynamic characteristics of video popularity distribution and wireless channels environment, with focus on reducing system energy consumption in the long term. The purpose of the proposed deep Q-learning optimization algorithm is aims to achieve optimal solutions for video caching and user association.

The emergence of 5G networks, especially mmWave 5G, enable bandwidth-intensive applications like 4K/8K and 360-degree video streaming. However, the perceived QoE of advanced ABR algorithms, including those based on reinforcement learning, remains inadequate on commercial mmWave 5G networks due to factors such as variable throughput and signal blocking. *COREL* [147], a primal-dual actor-critic RL approach, enhances existing methods by incorporating an additional critic network to estimate stall time and optimally adjust dual variables to minimize it.

NeuSaver [148] reduces energy usage in mobile by applying an adaptive frame rate to each video chunk without compromising QoE. In NeuSaver, the optimal policy uses the RL model to determine the appropriate frame rate for each video chunk. The RL model is efficiently and robustly trained using the Asynchronous Advantage Actor-Critic (A3C) algorithm, which is used to train deep neural networks to make decisions in varying environments. Based on previous observations, the RL model automatically learns the policy that maximizes the QoE objectives.

SACCT [149] is an energy-aware ABR streaming scheme designed for wireless edge networks. This scheme leverages a joint uplink transmission and edge transcoding algorithm that aims to maximize QoE while minimizing the energy consumption of the video player. To this end, the problem is formulated as a Markov decision process, and a DRL-based framework is applied to determine the optimal encoding bitrate, uploading power, edge transcoding bitrates, and frequency. SACCT integrates model-based optimization for continuous intra-frame resource allocation decisions and model-free deep reinforcement learning for discrete inter-frame bitrate adaptation decisions. This integration results in significant improvements compared with state-of-the-art approaches.

Brightness Scaling (BS) is a promising technique for enhancing energy efficiency in mobile video streaming. However, existing BS approaches ignore how the BS factor, video bitrate, and media context interact, leading to a mismatch between energy consumption and QoE. A user perception-based video Experience Optimization (PEO) model for energy-constrained mobile video streaming is presented in [150]. The proposed models jointly consider the intrinsic relationships among device motion state, BS influence, video bitrate, and user-perceived quality. Accordingly, it formulate motion-aware video streaming and user perception as an optimization problem, where the presented optimal algorithm maximizes the object function and thus proposes an online bitrate selection algorithm. As a results, PEO can improve the perceived quality up to 41% and save energy consumption up to 25%.

Adaptive video streaming algorithms usually prioritize video quality and ignore the impact of energy in their decisions which is critical in mobile devices. The energy-aware ABR algorithm in *E-WISH* [151] optimizes decisions by considering factors such as network throughput, client buffer fullness, video representation, and energy consumption. This approach aims to select the optimal representation for the next video segment, balancing quality and energy efficiency. According to experimental results, E-WISH can reduce the energy consumption in clients by up to 12% while improving the QoE by up to 52% compared to conventional approaches.

ABR applications are known for their high power consumption requirements. Built-in QoE optimization frameworks in modern mobile apps naturally lend themselves to integrated, proactive energy-aware app adaptation. However, this approach has two drawbacks; first, the application needs to predict energy consumption for each adaptation; second, it needs to incorporate the energy budget into its QoE optimization logic. In light of these challenges, proactive energy-aware adaptive streaming in mobile clients is explored in [152]. This study address two key issues: i) predicting power consumption when retrieving future video chunks in various candidate formats. ii) integrating energy constraints into the MPC algorithm to optimize QoE for forthcoming video chunks in an energy-aware ABR controller. The study showed that proactive energy-aware application, which integrates a user-specified energy drain budget into the QoE optimizer, can reduce application fluctuation and improve QoE over traditional reactive energy-aware application adaptation.

C. SUMMARY

Table 7 provides a comprehensive overview of the discussed papers focused on emerging applications and energy-aware technologies in video streaming.

VII. EMERGING NETWORK ARCHITECTURE AND QoE MANAGEMENT

A. WIRELESS NETWORKS AND QoE MANAGEMENT

Mobile communication has become essential in modern society due to its ubiquitous connectivity, accessibility, social integration, and emergency response capabilities. Consequently, mobile multimedia has emerged as a valuable service in next-generation heterogeneous networks. With the increasing use of 5G networks and the popularity of multimedia services, providing QoE over cellular networks will significantly impact both media content providers and network infrastructure. Therefore, to enhance customer satisfaction, QoE-aware video stream classification is essential. In this context, the authors in [153] introduced a framework that leverages machine learning and Enhanced Hyperparameter Tuning (EHPT) to optimize video streaming classification in 5G networks. The careful selection of parameter values through EHPT is crucial to the success. The performance of the ML approach is evaluated by considering the metrics of accuracy, precision, recall, and computation time to highlight the effectiveness of the classifiers in video stream classification. The simulation results indicate that the enhanced ML-EHPT ensemble classifier computes faster and displays greater precision, accuracy, and recall scores than the HPT decision tree classifier.

The application of reinforcement learning in MEC seeks to minimize the costs associated with valuable bandwidth in satellite networks. The approach proposed in [154] takes into account variations in network conditions within the Radio Access Network (RAN) and the heterogeneous computational capacities of edge servers. This model leverages the computational capacity of the MEC server along with a Super-Resolution (SR) algorithm to transmit high-quality video while conserving bandwidth on the satellite backhaul. In [155], the authors introduced a regression algorithm, Least Absolute Shrinkage and Selection Operator (LASSO), which is employed to train the dataset incorporating video, network, and client characteristics. Information obtained from various resources within wireless networks is utilized to develop models that predict future end-user QoE more accurately.

Since watching videos incurs mobile data charges for users, the traffic consumed by video downloads should be tailored to the user's specified limits. Considering this fact, the ABR algorithm proposed in [156] aims to improve mobile users' QoE during display time. To achieve this, the authors first analyzed real-world traffic and identified KPIs to formulate the QoE model. Then, introduced a Traffic-Aware Rate Adaptation strategy (TARA). Finally, a RL approach is proposed for the joint optimization of TARA, which aims to maximize the designed QoE KPIs and ensure the expected viewing experience for users. TARA provides a robust ABR process with traffic restrictions by selecting bitrates based on real-time user information such as network conditions, buffer status, and traffic amount. Performance tests conducted with both simulated and real-world network traces demonstrated TARA's sensitivity to environmental changes and its stability in dynamic bitrate adaptation.

In [157], a supervised ML classifier algorithm is trained to predict user satisfaction in DASH streaming over 5G network, considering network metrics, such as RTT, throughput, and number of packets. The proposed model operates without relying on application knowledge or specific traffic information. Similarly, SMASH [158] is a simple and lightweight client-side supervised ML approach for HAS. SMASH ensures a high QoE while maintaining consistent

TABLE 7. Summary of emerging application and energy aware QoE management schema. Each entry highlights key contributions, methodologies, and findings related to enhancing QoE.

Ref.	Description	ML	Edge/ Mobile	Energy-aware
[114]	Enhancing QoE with DNN by analyzing sequential features to predict optimal bitrates	✓	✗	✗
[115]	Enhancing low-bitrate video segments into high-quality streams with reinforcement learning	✓	✗	✗
[116]	An RL-based QoE enhancement employs a reward that jointly considers video quality & buffer status	✓	✗	✗
[117]	QoE enhancement through reinforcement learning in a VNF-assisted edge environment	✓	✓	✗
[118]	Improving QoE at the shared link by employing NN & RL to recommend video chunks for clients	✓	✓	✗
[119]	Enhancing QoE with deep learning, leveraging its potential for classification and regression tasks	✓	✗	✗
[120]	Fair QoE via server training algorithm to enhance action quality and quantity for client guidance	✓	✗	✗
[121]	QoE enhancement with a client request management solution uses DRL to handle concurrent user	✓	✓	✗
[122]	QoE enhancement with deep reinforcement learning without any assumption about environment	✓	✗	✗
[123]	QoE optimization in IoT streaming with jointly DRL and advanced actor-learner architecture	✓	✗	✗
[124]	Optimize bandwidth distribution and maximize QoE & fairness objectives in video streaming rewards	✓	✗	✗
[125]	RL-based model determining upcoming chunk transmission time in real environment	✓	✗	✗
[126]	ABR algorithm to predict the influence on metrics after taking different actions	✓	✗	✗
[127]	Implements online QoE perception schemes and a flexible adaptation algorithm for QoE diversity	✓	✓	✗
[128]	Monitoring user behavior and improving QoE with fair bandwidth allocation	✗	✓	✗
[129]	Tracing user behavior and build a metric to estimates QoE for a specific aspect of user actions	✓	✗	✗
[130]	Estimate playback behavior based on uplink request information to predict QoE parameters	✓	✗	✗
[131]	Improving the stable quality of video streaming DASH with assisted of DRL	✓	✗	✗
[132]	An pursue progress framework with data collection, meta-training, meta-testing, fine-tuning	✓	✗	✗
[133]	Quality-aware adaptive bitrate algorithm that selects a higher quality dynamic chunks	✓	✗	✗
[134]	Uses RL to develop high-quality ABR algorithms for Facebook's video platform	✓	✗	✗
[135]	Predicting VMAF scores for high dynamic range with supervised machine learning method	✓	✗	✗
[136]	Server-side learning-based scheme that offers QoE-aware bitrate guidance to clients	✓	✗	✗
[137]	QoE-aware bitrate guidance that complements the client-side heuristic-based ABR schemes	✗	✗	✗
[138]	A lightweight algorithm is proposed to remove redundant representations for each coded	✗	✗	✓
[139]	Saving energy in smart device during moving without decreasing video quality	✗	✓	✓
[140]	A DRL model adapts to dynamic conditions, excelling in diverse network scenarios	✓	✓	✓
[141]	An energy-aware DASH wrapper, minimizing energy usage while maintaining mobile user QoE	✓	✓	✓
[142]	Using DRL to ensures that data download attempts are synchronized with network conditions	✓	✓	✓
[143]	Neural ABR streaming with multipath transmission with tread-off between Qos and QoE	✓	✓	✗
[144]	Analyzing how path capacity (MPTCP) and data retransmission affect energy consumption.	✓	✗	✓
[145]	Energy & QoE-aware video streaming on mobile devices, prioritizing CPU scaling	✗	✓	✓
[146]	Deep learning-assisted optimization for video caching and user association in mobile network	✓	✓	✓
[148]	Energy-aware RL model determines the optimal frame rate policy for each chunk to achieve QoE	✓	✓	✓
[149]	Leverages edge transcoding algorithm to maximize QoE while minimizing the energy usage	✓	✓	✓
[150]	Energy optimization considering motion-aware video streaming, BS factors, and user perception QoE	✗	✓	✓
[151]	Energy-aware ABR algorithm considering throughput, buffer, quality, and energy costs	✗	✓	✓
[152]	Demonstrates that energy-aware app adaptation significantly reduces oscillation and enhances QoE	✗	✓	✓

performance across extensive streaming scenarios. This predictive model determines the best choice for the next video segment, regardless of streaming context complexity.

In [159] the performance of MEC-assisted and client-side adaptation methods was investigated in a multi-client cellular environment, focusing on QoE parameters, bandwidth utilization, and fairness. The results indicate that MEC-assisted algorithms generally improve fairness and bandwidth utilization compared to client-based algorithms across various settings. In addition, buffer-based algorithms can achieve significant QoE; however, they often perform less effectively than throughput-based algorithms in maintaining stable playback buffers during rapid fluctuations in bandwidth. The study concludes that the preparation of representation sets, along with factors such as playback buffer size and segment duration, significantly influence algorithm performance.

Deep Q-Network Reinforcement learning (DQNReg) [160] enhances classical deep Q-learning method, to improve video adaptation by leveraging experiences from exploring the wireless and 5G network environment. A QoE-based reward function is developed to ensure the learning strategy converges toward maximizing QoE outcomes. Additionally, research [161] address fair bandwidth allocation for video users in wireless networks, where bandwidth limitation impacts preserved video quality. This study introduces an ensemble learning model based on user behavior factors, which collaboratively allocates bandwidth for multiple users while leveraging the advantages of the SDN controller to optimize the user perception based on the user behavior.

Decentralized optimization for multicast ABR streaming in cache-assisted mobile networks is investigated in [162]. The study outlines ABR streaming in mobile networks as a Multi-source Multicast Multi-rate Problem (MMMP).

It presents a theoretical analysis of the computational complexity, convergence, and time-varying adaptation of distributed delivery algorithms independent of cache placement. While this model performs well in theoretical scenarios, it has not been applied in real-world wireless environments.

Congestion aware ABR streaming system over SDN enabled Wi-Fi network is introduced in [163]. In this architecture, an SDN-enabled Access Point (AP) collects client's KPIs as well as the Wi-Fi network conditions from the AP through OpenFlow. In case of congestion, it determines the segment bitrate for each client to provide seamless and high-quality streaming services.

B. EDGE COMPUTING AND QoE MANAGEMENT

While research on ABR streaming has advanced, further efforts are needed to validate both subjective and objective QoE performance models. On the subjective network parameters, some researchers have only considered the bandwidth parameter to evaluate the QoE of the ABR stream. While others include QoS parameters such as delay, packet loss, and jitter. The move towards 5G and the rapid increase in the number of end-users have demanded a higher quality of video streaming, motivating the development of new techniques. Some of such research involves the optimization of multimedia streaming frameworks while others focus on edge computing.

Moving toward edge computing offers low latency, bandwidth efficiency, scalability, and improved reliability. Resource-rich edge nodes bring storage and processing power closer to end-users, where valuable detailed information about traffic patterns, client behavior, and client distribution, helps better decision and service delivery. In this context, the Edge Computing Assisted Adaptation Scheme with ML (ECAS-ML) [164] is utilized for QoE management in HAS streaming. The proposed adaptation schema focuses on achieving the best QoE by managing the trade-off among bitrate, quality conciliation, and stalls. An automated model for QoE evaluation of ABR streams over wireless networks was introduced in [165]. This model assesses human factors (e.g., startup delay, quality fluctuation) that affect the QoE in wireless networks.

High user concurrency during peak live streaming periods leads to increased traffic at the edge. To mitigate this and enhance QoE, the authors in [166] utilize emerging edge computing technologies based on DRL algorithms. This approach smooths live streaming traffic by dynamically adjusting client bitrates and caching content at edge servers. Proposed QoE-aware Adaptive Video bitrate Aggregation (QAVA) scheme for HLS streaming, based on smart edge computing, aggregates all traffic requested by clients for the same live streaming and sets their bitrates based on network conditions, client states, and video characteristics. Compared to state-of-the-art bitrate adaption algorithms, QAVA schema improves fairness and achieves higher average client QoE levels.

An adaptive streaming schema introduced in [167] for edge computing environments aims to optimize the performance of multi-client adaptive streaming. The optimization is achieved by training the neural network model using the RL algorithm, which manage multiple clients under the supervision of a mobile edge. Service providers consider client-side observation, facilitated by edge computing, to develop policies that enhance QoE in multi-user environments. As a result, the proposed scheme improves overall perceived QoE and fairness.

When two clients are connected to the same access point, they may have different buffer fullness. Thus, equal priority to download video segment may result in buffer drain in one while the other has filled buffer. Currently most access networks treat all packets uniformly, lacking the flexibility to prioritize clients based on their immediate service needs. The widespread adoption of SDN has enabled the prototyping of agile control policies within access networks. Such a software reconfiguration capability can measure the impact on the application (e.g., QoE) and adaptively select a new configuration. *QFlow* [168] is a learning-based edge network configuration platform designed to address this problem. In the context of QFlow, a priority queue determines which clients should be assigned to each queue during each decision cycle. The results demonstrate that this approach enhances QoE by effectively prioritizing the appropriate clients in high-load scenarios.

To optimize the performance of video streaming services, quality adaptation should be intelligent and resilient to fluctuations in network conditions, while also minimizing bandwidth waste on the client side. The edge-assisted RL-based study in [169] aims to develop a robust adaptation policy that effectively responds to rapidly changing network conditions. The proposed neural network scheme learns client and network information, maximizing fairness among clients while efficiently utilizing available network bandwidth.

Some studies investigate HAS solutions that are capable of improving QoE in mobile networks characterized by rapidly changing conditions. The proposed QoE-aware HAS scheme [170] integrates a bandwidth prediction based on Gated Recurrent Unit (GRU) neural networks with an adaptive bitrate selection strategy to determine the suitable bitrate for each video chunk. The performance of the proposed solution has been validated through numerical simulations, which indicate that it is significantly more effective than the conventional methods.

The interaction of high bitrate video tends to aggravate network congestion and increase transmission energy, creating new challenges for wireless networks in terms of delivery delays and energy efficiency. To reduce the delay and energy consumption of video streaming over wireless networks, the proposed algorithm in *(JCCPA)* [171] jointly optimizes cache memory, computation, and power allocation with a Mixed Integer Nonlinear Programming (MINLP) approach. To end this, the main problem is divided into two sub-problems; power allocation, caching placement with computing decision problem. Additionally, the joint caching placement and computing decision problem is further divided into two sub-problems, which are solved iteratively.

VISCA [172] is an edge-based ABR algorithm that optimizes bitrate and video chunk sourcing by jointly considering network conditions, QoE objectives, and edge resource availability. It utilizes super-resolution techniques specifically trained on the most popular videos to enhance cached low-quality video at the edge, achieving significant quality improvements with minimal computational cost. Furthermore, the performance of super-resolution models is influenced by the caching strategy, as they can only be applied if the low-resolution video chunks are already cached. Hence, an new efficient caching strategy also applied for caching low-resolution chunk.

C. CDN NETWORKS AND QoE MANAGEMENT

A geographically distributed system like CDN has a significant impact on adaptive streaming. It enables efficient distribution of media content across edge points, ensuring high-quality delivery in the most efficient way for each individual device running the media player. Streaming via CDN, with minimized buffering and latency, enables users to achieve optimal video quality and resolution. CDNs accelerate streaming by placing content closer to end-users, thereby reducing latency, maximizing performance, and ensuring rapid downloading of media content. Whether in the form of private or shared caches, CDNs efficiently serve multiple users. With leveraging caching, frequently accessed resources are reused, which decreases network traffic and latency. The decentralized nature of a CDN increases the speed and efficiency of delivering content to end-users. However, performance may vary depending on regions and the time of day. Employing a multi-CDN strategy enhances reliability, flexibility, and cost efficiency [173]. In addition, streaming via a CDN enhances security by mitigating Distributed Denial-of-Service (DDoS) attacks. Redundant CDNs with multiple access points reduce the risk of a single point of failure.

The main potential benefit of CDN architecture relies on cache efficiency. CDNs architectures are typically organized as a cache hierarchies. When a user makes a request, it is directed to a CDN edge server. If the requested object (video segment or chunk) is already cached there, the edge server responds directly to the request. Alternatively, if the object is not cached, the edge server can forward the request to a parent CDN (a higher-tier server in the hierarchy) and ultimately to the origin media server. Once the object is fetched, it is then transferred to the client. The authors in [174] introduced a multi-tier caching schema that considers various system design factors, such as the limited caching space at CDN sites, allocation of CDN resources for video requests, bandwidth allocation, edge-cache capacity, and caching policies. Their focus is on minimizing a performance metric known as Stall Duration Tail Probability (SDTP). Evaluations of prominent CDNs like Akamai, Fastly, and CloudFront indicate that integrating CDN awareness into ABR algorithms can significantly improve throughput prediction accuracy and enhance overall ABR algorithm performance.

Study [175] contends that the edge caching environment is highly dynamic, characterized by diverse request patterns. Hence, this complexity suggests that current rule-based and model-based caching solutions may not adequately address the challenges presented by such dynamic edge environments. Furthermore, while collaborative caching has been proposed to optimize limited storage across individual edge servers, the request similarity among neighboring edges can be highly dynamic and varied. This diversity can potentially undermine the benefits of traditional cooperative storage systems designed for CDN environments. They proposed DRL approach for intelligent video caching at the edge, aiming to minimize both the access latency and traffic cost. Furthermore, the study shows that content access patterns and content similarities among different edge regions are highly heterogeneous and dynamic in both temporal and geographic dimensions.

Video startup time significantly influences perceived QoE. Legacy approaches estimate the network throughput for each user, which can incur a significant overhead. *Rldish* [117] is a RL-based schema deployed on the edge CDN server, solves this problem by dynamically selecting a suitable video segment for new live client. As an NFV function on a real HTTP cache server, Rldish collects real-time QoE observations from the edge without requiring client-side assistance. It then uses these KPIs as real-time rewards in its reinforcement learning framework. As a results, it improves legacy initial video selection process and enhances average QoE.

Predictive CDN selection discussed in [176], where the proposed meted aims to provides cost-effective video delivery by taking into account clients update information while displaying video. Predictions are used by the media server to take action when the player requests an MPD update. The media server can decide to keep the same MPD or change it to another available CDN from which the content can be downloaded. However, this method has some disadvantages. Firstly the media server needs to prepare a different MDP for each CDN. Moreover, monitoring the client activity from the media server increases the media server load, and ultimately results in slow response.

The feasibility of CDN switching for long video session and a privacy-preserving CDN-ISP collaboration framework using SDN for VoD streaming is highlighted in [177]. In this framework, the video provider gives the client a brief list of CDN addresses along with the ISP server interface for connection. The client establishes a secondary connection to this ISP server. The ISP server continuously queries the short list of CDN addresses to identify the CDN with the best network availability. However, in practice, OTT providers typically collaborate directly with CDNs, which are the entities that maintain relationships with ISPs. Moreover, to keep network latency under control, it is not feasible to constantly increase the number of CDN servers in different geographical locations on-demand.

A Peer-to-Peer (P2P) CDN serves as a sophisticated alternative to complement traditional CDN servers, utilizing the advantages of both P2P and CDN technologies. The hybrid P2P-CDN architecture for live video streaming, which leverage edge computing and NFV, is presented in [178]. This study formulates the action decision problem as a Mixed-Integer Linear Programming (MILP) optimization model. Additionally, an online learning approach based on unsupervised machine learning is introduced to mitigate the time complexity associated with the optimization model.

P2P-CDN live streaming systems are being deployed to reduce the stream delivery overhead of CDNs. However, the inherent fluctuations and heterogeneity in peers' upload contributions make it difficult to maintain QoS in these systems. To deal with this challenge, proposed Serviceability-aware Overlay Management Strategy (SOMS) for CDN-P2P live streaming systems in [179], organizes peers based on their serviceability, which is defined by factors such as stability, availability of stream chunks, and upload/download capacities. Peers with higher upload capacity are part of an extended CDN tree to facilitate stable seeders. A portion of these capacities is allocated to create virtual sources, ensuring quick startup for seamless playback. The topology adapts dynamically, allowing peers adjust partners based on serviceability and upload capacity to maintain QoS during churn.

The shortcomings of existing hybrid P2P-CDN video streaming solutions and the potential of emerging multi-access edge data centers are discussed in [180], and a novel SDN hosted P2P-CDN service architecture is introduced. A key feature of the proposed service architecture is that both CDN access by peers and P2P video streaming between peers within edge access networks are fully controlled through the collaboration of OTT and network service providers to optimize video KPIs. As a result, managing all these activities at edge data centers reduces the load on CDN servers while overcoming QoE fluctuations per flow and unfairness among multiple heterogeneous video-resolution clients on dedicated access network slices.

The dynamic joining and leaving of peers (churn) is a common phenomenon in P2P systems, where there is no assurance that a specific peer will remain connected throughout the entire video streaming session. Moreover, in a P2P-CDN architecture, there is no guarantee that each peer contains all the representations of a single video file. Finally, the mere presence or absence of certain video chunk/bitrates cannot be considered the sole criterion for the decision making of participating peers. The behavior of the underlying network significantly impacts video streaming performance, as it's possible that the peer containing the desired chunk/bitrate may not be reachable via the most optimal route.

D. QUICK UDP INTERNET CONNECTIONS

QUIC has recently been deployed to improve the performance of video streaming. It offers advantages over TCP, such as security, reduced latency, improved congestion control, and faster connection establishment. These features make QUIC particularly useful for delivering smooth and high quality video streaming experiences over the Internet. Adaptive Stream-scheduler Multipath QUIC (ASMQ) [181] is a scalable video coding framework using multipath QUIC. The proposed framework has been used in DASH-SVC to address network congestion caused by sudden increase in data traffic due to mismatches between flows with different priorities and paths over multipath QUIC. ASMQ schedules prioritized streams onto paths with different qualities. Additionally, ASMQ's server-client feedback mechanism can adapt to network congestion conditions in real-time.

In terms of content delivery and security, network providers face a serious challenge in managing their networks due to the deployment of end-to-end security protocols such as HTTPS and QUIC. They require clear visibility of their networks' traffic to monitor and manage QoS/QoE impairments in video streaming services in the most effective and efficient manner. To bridge this gap, proposed ML based solution in [182] leverages a random forest classifier for a better QoE inference accuracy. The proposed solution utilizes network and transport layer information to infer QoE factors. This enables network providers to promptly react in real-time to any disruptions in QoE for encrypted video traffic. As a result, it facilitates effective monitoring and management of video streaming services, ensuring a seamless and reliable display experience for end-users.

The study in [183] argues that in indoor scenarios where QUIC is implemented, overall performance surpasses that of legacy network protocols. These findings are encouraging as they demonstrate that by appropriately adjusting the QUIC network protocol can achieve desired levels of QoS and QoE for interactive multimedia content. Considering the requirement of video streaming applications that prioritize low latency and high video quality, a DRL based multipath QUIC (MPQUIC) scheduler proposed in [184] aims to improve video streaming quality. Although the evaluation results show that MPQUIC outperforms the legacy schedulers, user experience may be degraded by choosing the wrong action at the initial stage, due to the nature of reinforcement learning. *MAppLE* [185] is a MPQUIC latency evaluation platform that provides the instrumentation needed to evaluate and develop MPQUIC stream multiplexers, stream schedulers, and multipath packet schedulers. According to the experimental results, in a lossy asymmetric heterogeneous wireless network, the proposed scheduling reduces outlier delay compared to existing scheduling algorithms and improves QoE in video streaming by minimizing rebuffering.

QUIC promises a more responsive and secure web experience, with design principles aimed at eliminating the head-of-line blocking problem, introducing fast connection facilities and integrating transport layer security. However, the authors in [186] found that QUIC is not sufficiently fast on high-speed Internet and adversely affects the video streaming applications with a reduction of up to 9.8% in video bitrate. The authors in [187] show that ABR techniques are sensitive to sudden changes in client-perceived bandwidth, making them better suited for TCP than QUIC. The multiplexing of audio and video by QUIC over UDP introduces a latency for audio segments that is not accounted for in throughput calculations. As a result, ABR algorithms may incorrectly choose bitrates over QUIC connections. This suggests the need for research into ABR techniques optimized specifically for QUIC. It is highlighted in [188] that QUIC does not offer advantages in networks with typical loss and delays. However, in networks with long delays, QUIC provides higher playback bitrates and lower re-buffering rates. Hence, QUIC's benefits are more pronounced in networks with significant delays and moderate loss. This makes QUIC particularly beneficial for improving user experience in regions where early-generation 3G networks are prevalent.

Integrating DASH with the QUIC partial reliability concept to reduce playout interruptions and increase the quality of immersive content delivered over lossy networks is discussed in [189]. In this approach, the DASH scheme makes quality and prioritization decisions based on changing network conditions and user visibility, respectively. By using information from the user's viewport to prioritize data, the deployed framework deliver content within the viewport at the highest possible quality (based on the available bandwidth) reliably, while delivering the rest unreliably. The results indicate that under ideal network conditions, HTTP/2 is preferred over other protocols. However, in the presence of packet loss, reliable delivery mechanisms become crucial. While HTTP/2 shows high intolerance to loss, reliable QUIC has demonstrated some improvement but still lacks robustness. Practical coding in a low latency transport protocol is integrated in robust QUIC (rQUIC) [190]. Integrating rQUIC with DASH enhances perceived video quality for end-users and enables transmission of higher resolution frames in scenarios where video playback quality is already satisfactory.

E. SUMMARY

Table 8 summarizes all the discussed papers regarding network architecture and QoE management.

VIII. VIRTUALIZATION AND SOFTWARIZATION TECHNOLOGY

A. NETWORK FUNCTION VIRTUALIZATION

Nevertheless, while CDNs and adaptive streaming offer significant flexibility, legacy network architectures based on the *TCP/IP* protocol are no longer efficient. Designing a new network protocol and architecture is more costly and, in practice, it is impossible. Network configuration is the other concern that should be considered. The heterogeneous nature of the technologies, infrastructures, and applications has complicated the management of such systems. While network condition is more ad hoc, it is hard to configure the network based on predicted policy. In addition, any policy changes or faults require the reconfiguration of network devices.

The current state-of-the-art network configuration is more time-consuming. Vertical integration in the current IP network enforcing automatic configuration and preventing innovation acceleration (e.g., moving to IPv6) in network infrastructure [191]. As an example, due to the stationary of the current IP network, it takes more than five years to fully design, evaluate, and deploy a new routing algorithm [191]. Finally, a clean-state approach to completely redesign the ongoing network is ultimately more expensive and impractical in real-world scenarios. To address this problem we need network transforming. Results of the most agile, flexible, and programmable network infrastructure are better aligned with the needs of application workloads.

The NFV-SDN presents abstractions of the network virtualization that address various networking challenges without touching the underlying infrastructures. In terms of network performance, traditional implementation based on the tightly coupled infrastructure are no longer effective. Network functions or middle-box (e.g., cache, load balancing, IDSs) are used to modify workload (packets and flows) in sophisticated ways [192]. However, using network functions as a temporary solution to enhance network flexibility can introduce new management challenges. Virtualization is based on the fact that virtual machines installed on-demand and share the same hardware resources, while the network should assign identical attribute to the computing layer. Further, on-demand nature of construct, emigrate and destroy are the momentous of virtual machines.

The individual concept of NFV has a successful experience in *MPLS* (virtualized path) [193], *NAT* (Virtualized IP address), and *VLANs* (Virtualized L2) [194] domains. However, in the absence of global configuration, the distributed nature of virtualized functions increases both configuration costs and complexity over time. More importantly, vendor-specific NFV solutions contract with NFV original concept [9]. With the SDN assistance, distributed control and configuration can be centralized in a single point. As a consequence, the SDN platform should be open to applications and hardware from multiple vendors. Additionally, in terms of scalability and reliability, control redundancy should be considered.

The NFV orchestration which is in the charge of provisioning for virtualized network functions, controlled by the SDN controller through southbound interfaces [195]. The SDN controller computes the optimal function assignments (e.g., assigning virtual cache functions to certain location or node) and translates the the logic policy characteristic into optimized routing paths. Function assignments are enforced by the NFV orchestration, while the controller steer the traffic by installing forwarding rules into forwarding devices. NFV leverage virtualization to separate software instance from

TABLE 8. Summary of network architecture and QoE management schema.

Ref.	Description	ML	Edge/ Mobile	NFV-SDN	CDN	QUIC
[153]	Enhanced ML classifiers were developed for video streaming categorization	✓	✓	✗	✗	✗
[154]	Minimizing satellite communication bandwidth costs with super-resolution	✓	✓	✗	✗	✗
[155]	Train models to analyze the correlation between network, video, and client for QoE	✓	✓	✗	✗	✗
[156]	Present a traffic-aware rate adaptation strategy using centralized RL to optimize QoE	✓	✓	✗	✗	✗
[157]	Training a ML to predict user QoE during video sessions based on QoS metrics	✓	✓	✗	✗	✗
[158]	A supervised ABR predictive model to select the best next video segment	✓	✓	✗	✗	✗
[159]	Analysis of MEC-assisted and client-side ABR algorithms in a cellular network	✗	✓	✗	✗	✗
[160]	An RL approach creates a segment-wise QoE-based reward function to maximize QoE	✓	✓	✗	✗	✗
[161]	Ensemble learning uses an SDN to optimize user perception based on behavior	✓	✓	✓	✗	✗
[162]	Decentralized multicast ABR optimization in cache-assisted mobile networks	✓	✓	✗	✗	✗
[163]	Congestion aware ABR streaming system over SDN enabled Wi-Fi network	✗	✓	✓	✗	✗
[164]	Edge assisted adaptation scheme with ML algorithm for QoE management	✓	✓	✗	✗	✗
[165]	Considering human factors for improving QoE in wireless networks	✗	✓	✗	✗	✗
[166]	Utilizing DRL algorithm for mitigating edge traffic in peak time	✓	✓	✗	✗	✗
[167]	ML-based approach to optimize multi-client adaptive streaming	✓	✓	✗	✗	✗
[168]	Learning-based edge network configuration platform for client priority	✓	✓	✓	✗	✗
[169]	Edge-assisted RL schema provides robust adaptation and maximize fairness	✓	✓	✗	✗	✗
[170]	An ABR scheme uses bandwidth prediction and bitrate selection to enhance QoE	✓	✓	✗	✗	✗
[171]	QoE-aware and energy efficiency solution using nonlinear programming	✗	✓	✗	✗	✗
[172]	Utilizes super-resolution to enhance cached low-quality video at the edge	✗	✓	✗	✗	✗
[173]	Enhance reliability, flexibility and cost efficiency with Multi-CDN schema	✗	✓	✗	✓	✗
[174]	A video streaming model with a central server, CDN sites, and edge caches	✗	✗	✗	✓	✗
[175]	Enhancing QoE in multi-tier edge caching with DRL to minimize stall duration	✓	✓	✗	✓	✗
[176]	A solution that predicts bandwidth and latency for MPD updates and CDN switching	✗	✗	✗	✓	✗
[177]	CDN-ISP collaboration to switch CDNs and enhance QoE in VoD streaming	✗	✗	✗	✓	✗
[178]	Hybrid P2P-CDN schema that alleviates decision time complexity with online learning	✗	✗	✓	✓	✗
[179]	Maintain QoS in P2P system based on serviceability and upload capacity	✗	✗	✗	✓	✗
[180]	A P2P-CDN schema reduces CDN load and QoE fluctuations via chunk scheduling	✗	✓	✗	✓	✗
[181]	Prioritizing streams based on application information and evaluating multipath qualities	✗	✓	✗	✗	✓
[182]	Uses network and transport layer data to infer QoE factors like startup delay and stalls	✗	✗	✗	✗	✓
[183]	Shows that QoS/QoE for interactive multimedia can be achieved with QUIC adjustment	✗	✗	✗	✗	✓
[184]	RL-based MP-QUIC scheduler using delay and throughput to minimize download time	✓	✗	✗	✗	✓
[185]	A multipath scheduling with redundancy for higher throughput and rapid loss recovery	✗	✗	✗	✗	✓
[186]	Examines QUIC's performance over high-speed networks and shows it's not fast	✗	✗	✗	✗	✓
[187]	Shows ABR with client-perceived bandwidth is better suited for TCP than QUIC	✗	✗	✗	✗	✓
[188]	Shows that QUIC does not offer advantages in networks with loss and delays	✗	✗	✗	✗	✓
[189]	Integrating DASH with the QUIC to increase the quality over lossy networks	✗	✗	✗	✗	✓
[190]	Integrating DASH with robust QUIC to enhances perceived video quality	✗	✗	✗	✗	✓

hardware platform and gives more agility to network services by decoupling functionality from location [196]. However, by taking advantages of NFV-SDN, it is important to ensure that network performance is remains at least as good as previous and does not bring new challenges such as security risk. In addition, VNFs can be moved between different platforms. However, movement should not negatively impact QoS parameters such as packet loss and delay.

B. SOFTWARE DEFINED NETWORKING

In addition to efficiency of video streaming technologies, the performance of these applications is closely tied to underlying network conditions. Providing better network capacity helps to increase the quality of video delivered to clients. At first glance, tightly coupling control and data plans results in network resiliency. However, this relatively static architecture reduce network flexibility and eliminate easier innovation. Creating and introducing a new abstraction of networking with respect of simplifying management is achievable by breaking network control problem into traceable pieces [191]. Therefore, SDN technology can be a good alternative for network operators offering services to video streaming providers [13]. SDN is a preferred solution for high demand resources, unpredictable network traffic patterns, developed in response to future networks demands.

SDN breaks the vertically integrated network architecture by separating control logic (control plane) from the underlying network infrastructure (forwarding plane) into a logical central programmable model and gives more agility to network functions. In lack of SDN centralized control and management capabilities, current IP networks operators need to configure each device individually. The main concept of the SDN architecture realise on modifying heterogeneous forwarding devices to support data plane programmability via open and standard interfaces (e.g, OpenFlow), which is effortful in legacy network due to the distributed nature of

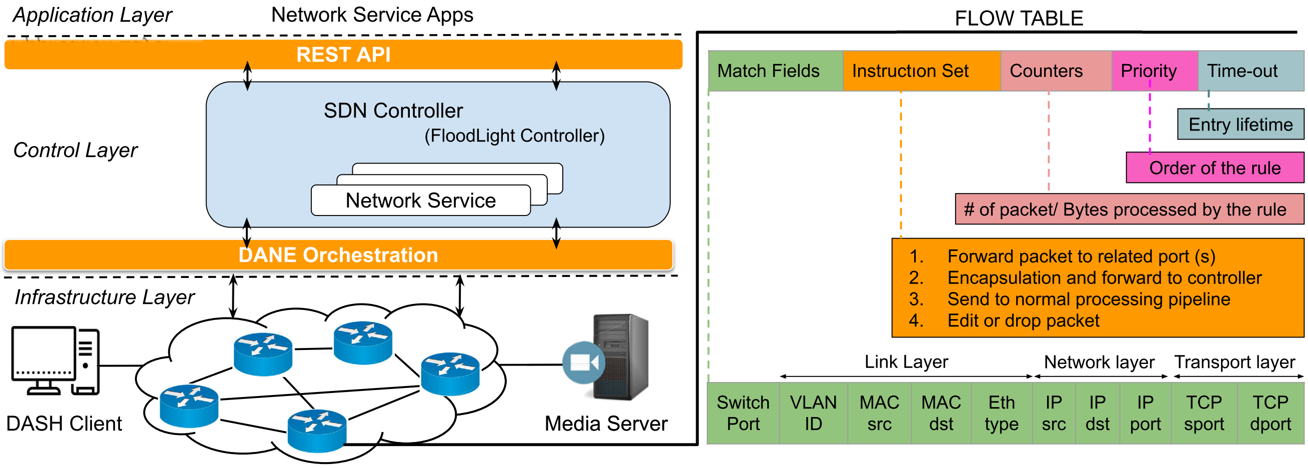


FIGURE 11. Illustration of the SDN architecture and forwarding device's flow table (right side). The central controller in the SDN network updates forwarding devices' routing tables with new routing information, enhancing their agility in packet routing.

the control plane. Network operators develop and implement application specific routing strategies by integrating network intelligence into the control plane. The main concept of the SDN architecture relies essentially on modifying forwarding devices to support flow tables. As shown in Fig. 11, top-down deceleration of the SDN architecture include five layers:

1) *Management plane:* This layer contains a list of applications used for traffic management (e.g., routing algorithm and load balancing), security (e.g., intrusions detection system), and network control protocol (e.g., SNMP). All applications can leverage almost same network information for policy management and effective decision making. The policies of these applications are ultimately executed by forwarding devices.
2) *Northbound Interface:* Northbound interfaces allows communication among the higher-level components such as application and network operating system, known as the *SDN Controller*, via *RESTs APIs*. Typical, the purpose of these interfaces is to abstract the inner-workings of the network and provide convenience for application developers.
3) *Network Operation System:* Controller, serves as the core of the SDN, loading the forwarding device's tables with application policies. Basic network functions (e.g., link manager, path manager, flow manager) assist the controller in collecting network metric and statistics. This abstraction of network information simplifies the programming of management applications.
4) *Southbound Interface:* Southbound *APIs* hide details of the individual network devices and facilitate efficient control over the network. This characteristic enables the controller, as well as application, to dynamically make changes according to real-time demands without struggle with low-level configuration.
5) *Data Plane:* Also known as forwarding plane, encompasses programmable physical equipment, such as OpenFlow switches, which have less intelligent about network. Forwarding devices contains one or more forwarding tables and well-defined instruction set (e.g, flow rules) used to take action on incoming packet (e.g, forward, drop, modify).

C. NFV AND SDN CORRELATION

Although the concept of SDN and NFV are independent, they are highly correlated in practice. The core similarity between them is their use of network abstraction and heavy reliance on virtualization. SDN and NFV are performing different roles, but leveraging NFV-SDN specifications enable networking architectures more flexible and result to network infrastructure agility. While SDN optimize network infrastructure by separating network control and forwarding function, NFV optimize development of the network functions. As shown in Fig. 12, NFV operates at the upper layer of the *OSI* model and provides basic networking functions, while SDN orchestrates and configures network functions for specific uses. SDN benefit the advantage of virtualization in order to provide programmable configurations and quickly optimize network resources. A variety of programmable interfaces permits network programming, making flow-base routing and dynamic traffic adjustment possible to meet fluctuating demands.

Despite NFV and SDN are not completely the same technology and each operating on different network layers, they are mutually beneficial. While NFV reduces the cost of network operational expenditure, SDN leverages NFV to enhance its performance by simplifying the compatibility with legacy network and making networks programmable. In addition, middle box, as a new and temporary network features designed for flexibility, introduce new challenges in network management. Fortunately, leveraging SDN technology (also know as *SDN Controller*) provides greater flexibility in network management while provides global view of the network.

Albeit NFV introduce a way to separate network services (e.g., caching, load balancing) from the underlying hardware by transferring network functions from dedicated hardware to general purpose software, such as hypervisor, SDN rely on NFV and virtual centralized controller. With the assistance of NFV, SDN can steer traffic dynamically based on user requirements. Thus, resource utilization can be achieved through the combination of NFV and SDN.

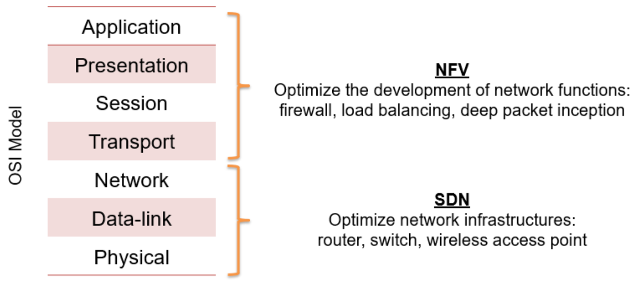


FIGURE 12. **OSI model and NFV-SDN correlation.**

D. EMERGING NETWORK ARCHITECTURE AND NFV-SDN PRINCIPLE

Dealing with fluctuating network traffic and uncertain underlying network dynamics, developing an effective orchestration model with low network costs remains a critical challenge in NFV-enabled networks. To address this, the authors in [197] proposed a DRL-based QoE-aware adaptive Online Orchestration (DRL-QOR) approach to adapt to network dynamism. The proposed models formulates stochastic resource optimization using a Parameterized Action Markov Decision Process (PAMDP), where QoS/QoE requirements are primary factors in shaping the reward function. The objective is to maximize QoE while ensuring compliance with QoS constraints.

Considering that various influencing factors (e.g., human, video content, and network factors) affect QoE perceived by end-users, the authors in [198] focused on improving QoE estimation by exploiting valuable human behavior, system factors, and context information. They proposed a QoE management approach based on supervised ML algorithms in SDN/multi-access edge computing, implementing a QoE prediction model to optimize the video delivery transmission chain. Similarly, the incorporation of SDN and MEC, assisted by a DRL algorithm, presents QoE optimization for 3D video streaming, as discussed in [199].

The concept of NFV-SDN is leveraged in *OSCAR* [200], where Virtual Reverse Proxy (VRP) and virtual Transcoder Function (VTF) are used to optimize resource utilization. At the edge, VRPs gather clients' requests and transmit them to an SDN controller. Subsequently, the SDN controller identifies an optimal set of multicast trees to stream the requested videos from a suitable origin server to the VRPs. OSCAR delivers only the highest requested quality directly from the origin server to an optimal group of VTFs via a multicast tree. These selected VTFs then transcode the video segments and multicast them to the VRPs that made the request. However, the transcoding process can introduce significant time complexity in practice.

The NFV-SDN concept as a virtual cache (vDANE) has been proposed in [201], [202], and [203]. This approach aims to enhance bandwidth utilization and improve perceived QoE by dynamically deploying caches on-demand in optimal locations based on client requests. As shown in Fig. 13, the SDN controller forwards client's request to appropriate cache that has already cached the requested chunk/bitrate. The feature of caching at the edge also studied in *EdgeDASH* [204], which is a network-assisted control logic running at a Wi-Fi access point to facilitate the use of cached video content. Although EdgeDASH effectively boosts cache hits and reduces buffer stalls by only changing the client request by a single quality level, it also introduces increase in session instability.

Energy efficient SDN-enabled ABR streaming at the mobile edge is discussed in [205]. To achieve this goal, the proposed system utilizes multipath technology to surpass the constraints of a single Wi-Fi access point. Additionally, it employs the Luby Transform (LT) code for forward error correction, ensuring flexible and dependable data transmission across multiple APs. The proposed system achieves real-time operation by finding a nearly optimal solution with low computational complexity. Experimental results demonstrate that the proposed system delivers high-quality video streaming services while maintaining a stable playback buffer and low energy consumption.

Despite significant improvements in QoS, network operators will still face significant challenges with increasing 5G video traffic volumes. As a result, the focus of network quality has recently shifted from network provider QoS to user-perceived QoE. It is assumed that 5G networks should be capable of delivering Ultra HD video streaming, and QoE-aware techniques must meet the quality standards expected by users. In study [206], a combination of SDN and network coding technique is explored to enhance adaptive video streaming in 5G network management and operation. SDN provides a comprehensive view of network resources and traffic, allowing for insights into optimal networks for implementing network coding. Experimental results indicates that video streaming using network coding technique outperforms the streaming without network coding. An NFV-SDN enabled architecture in 5G/6G networks for video streaming applications was introduced in *SARENA* [207]. The proposed architecture involves a set of Service Function Chains (SFCs) based on the QoE requirements of various MS services. Similar to a previous study, SARENA formulates the problem as a central scheduling optimization model, which is executed at the SDN controller.

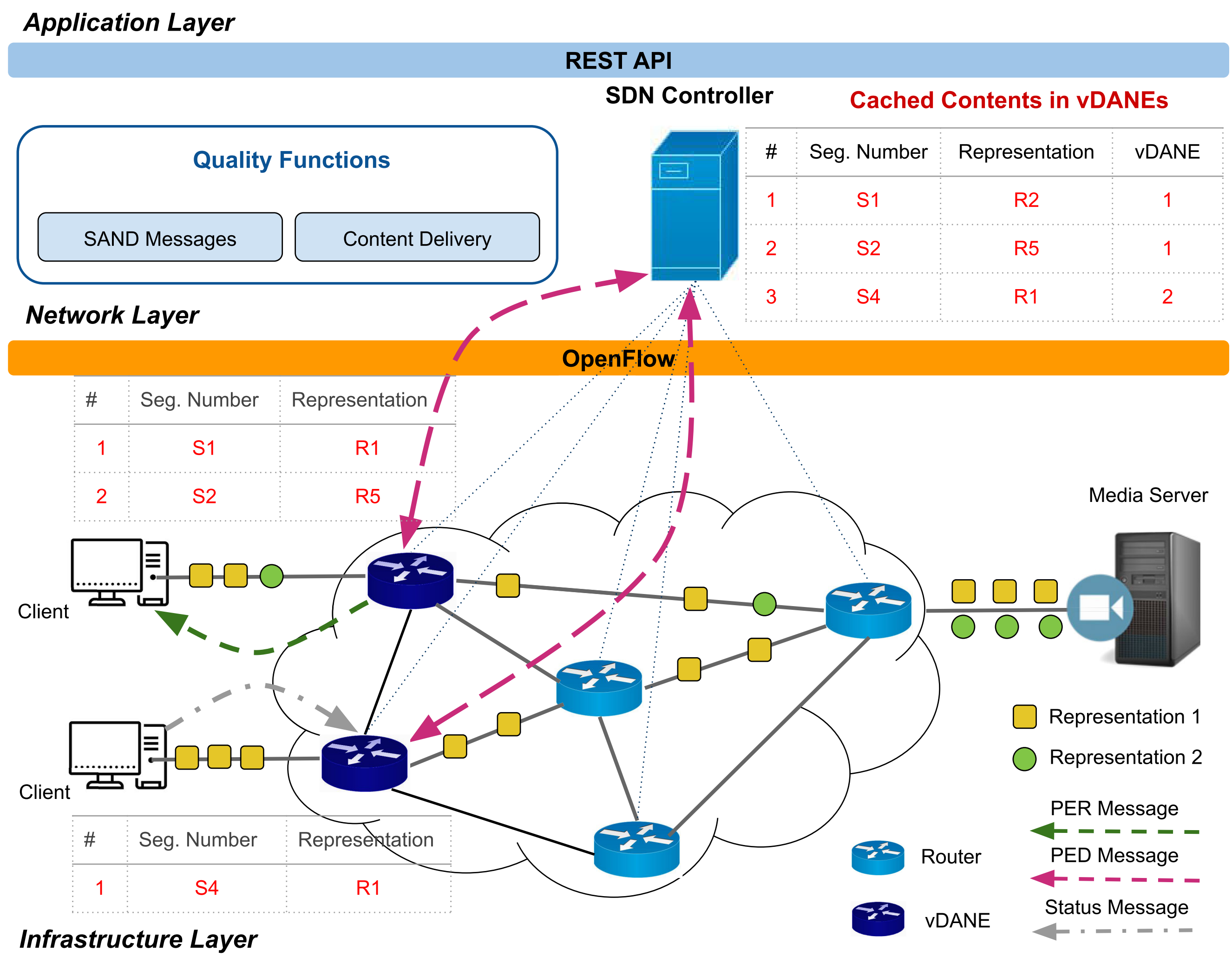


FIGURE 13. SDN-enabled QoE-driven ABR streaming schema, leveraging distributed NFV at the CDN-edge. The central controller has insight about network traffic patterns and available resources, enabling it to respond effectively to client requests.

The main objective of most QoS routing research studies is to develop routing algorithms that can provide the best paths to satisfy customer Service Level Agreement (SLA). Achieving this goal requires an updated global view of the network and the ability to predict the best routing paths. The SDN-based routing algorithm proposed in [208] seeks to identify the optimal path that simultaneously supports user SLAs and enhances perceived QoE stability. To achieve this, a flow placement strategy is implemented, designed to minimize the concentration of prioritized flows. This approach protect prioritized packets against losses during congestion by employing traditional QoS enforcement techniques, such as traffic shaping, priority queuing, or other relevant mechanisms.

The system architecture proposed in [209] leverages SDN technology to assist clients in choosing the optimal video codec and bitrate based on current network conditions, as well as directing video packets through suitable paths. In this architecture, layered video is employed, where each additional layer contributing to enhances quality. The controller estimates packet loss probability by considering video codec properties, layer bitrates, and network capacity. This estimation helps the controller choose suitable codec types and video qualities for clients, facilitating effective network management.

VQoSRR [210] is a QoS-based routing and resource reservation framework for ABR streaming. This QoE-aware SDN framework enables SDN networks to meet video requirements and improve user experience over best-effort networks. A developed routing algorithm installs routing path on forwarding devices, and redirect traffic to an alternative path if QoE is compromised. However, despite its effectiveness in improving QoE, the proposed framework suffers from scalability. *ARARAT* [211] is an edge-assisted collaborative framework for ABR streaming that utilizes NFV-SDN technology to enhance response time and reduce service cost. In order to deal with the high time complexity of the proposed centralized optimization model, the study

TABLE 9. Summary of virtualization approaches and QoE management schema.

Ref.	Description	ML	Edge/ Mobile	NFV-SDN	CDN
[197]	DRL approach dealing with fluctuating network traffic and uncertain underlying network	✓	✗	✓	✗
[198]	QoE estimation using human behavior, system factors, and contextual information	✓	✓	✓	✗
[199]	QoE optimization in 3D video, incorporating MEC/SDN with DRL algorithm	✓	✓	✓	✗
[200]	Optimizing resource utilization with virtual proxies and transcoding in NFV-SDN networks	✗	✗	✓	✗
[201]	Enhancing QoE with efficient virtual cache migration in NFV-SDN network	✗	✗	✓	✓
[202]	Enhancing QoE with utilizing virtual cache in NFV-SDN network	✗	✗	✓	✓
[203]	Enhancing QoE with leveraging distributed virtual cache at the edge	✗	✗	✓	✓
[204]	Introduce EdgeDASH is a network-side logic at a WiFi AP for optimizing cached video	✗	✓	✓	✗
[205]	Uses multi-path technology to overcome single Wi-Fi AP limitations in SDN networks	✗	✓	✗	✗
[206]	Fascinating video streaming in 5G networks combines SDN and network coding	✗	✓	✓	✗
[207]	Enhancing QoE for various mobile clients with NFV-SDN in 5G/6G networks	✗	✓	✓	✗
[208]	A routing algorithm that improves QoE stability while meeting service level agreements	✗	✗	✓	✗
[209]	Improving QoE by selecting the optimal video codec and efficient transmission path	✗	✗	✓	✓
[210]	QoS-based routing and resource reservation for ABR streaming in a QoE-aware SDN	✗	✗	✓	✗
[211]	QoE enhancement with edge assisted collaboration framework in NFV-SDN network	✗	✓	✓	✗
[212]	Using periodic and adaptive routing for optimal path selection to insure smooth adaptation	✗	✗	✓	✗
[213]	Combines SDN advantages with a UDP design to improve video efficiency and quality	✗	✗	✓	✗
[214]	Enhances video streaming quality by leveraging SDN advantages and SAND features	✗	✓	✓	✓
[215]	An optimization model to efficiently serve client requests with optimal cache server	✗	✗	✓	✓

introduces heuristic approaches. These heuristic generate near-optimal solutions through efficient cooperation between the SDN controller and edge servers.

To enhance service quality, the proposed schema in [212] employs different algorithms on both the client and network side. The client-side adaptation algorithm is designed to adjust based on buffer occupancy and network throughput, ensuring smooth adaptation without causing buffer stalls. On the network side, an SDN controller monitors and manages the network using two routing policies; periodic routing and adaptive routing. In the periodic routing method, bandwidth stability and throughput are considered to select the optimal path from the client to the server. Additionally, clients have the capability to actively request a new path that best meets their requirements. A SDN-driven reliable transmission architecture introduced in [213] for enhancing quality of ABR streaming. In this architecture, a retransmission mechanism within an SDN-switch that incorporates a buffering agent, mitigate packet loss. Overall, proposed architecture provides reliable and high quality video streaming, however it has difficulty in scalability and congestion control.

Determining the desired target quality for each client in SDN networks by leveraging the capabilities of MPEG-SAND provides an overview of the network status and available resources. In [214], the CDN-edge cache retrieve the requested segment from its neighbors; this is faster and more bandwidth efficient than requesting from the remote origin servers. Given this architecture and following the local hit and CDN hit steps, fetching from the origin server has the lowest priority. *ES-HAS* [215] is an edge and SDN-assisted ABR streaming framework designed to efficiently serves clients' requests in a time-slotted manner by selecting an optimal cache server. In case of a cache miss, a client's request is served either by an optimal replacement quality from a cache server or by the original requested quality level directly from the origin server.

E. SUMMARY

Table 9 summarizes all the discussed papers on virtualization approaches and QoE management.

IX. CONCLUSION

Video streaming has become a major use case for the Internet. The increasing popularity of mobile devices and emerging applications, such as 4K and 360-degree videos, require careful sharing of network resources among users to achieve QoE and fairness objectives. In recent years, QoE has gained significant research interest as an important factor in determining network efficiency. Evaluating and optimizing the video streaming ecosystem consisting of server agility, network QoS, and client adaptation algorithms leads to improved QoE. Although QoS metrics are not always directly linked to perceived quality, they significantly affect the performance of the delivery network, and consequently affect perceived QoE.

This study surveys various aspects of enhancing perceived video quality through both subjective and objective metrics. It also provides a comprehensive review of research published in the last five years, focusing on intelligent machine learning and network-assisted approaches. The use of advanced technologies such as NFV-SDN, CDN-Edge, machine learning and artificial intelligence offers significant benefits for increasing QoE. These improvements not only optimize network performance and reliability, but also facilitate content delivery. By leveraging these technologies, service providers can achieve higher levels of user satisfaction, improve network performance, and ultimately gain a competitive advantage in today's digital landscape.

Improving network efficiency plays a crucial role in enhancing user perceived quality in several key ways:

- *Reduce Latency*: Efficient networks reduce latency and improve response time, ensuring that users experience faster load times for content and applications. As a result, clients can download and display video segments at higher bitrates with minimal stall. This translates to a smoother and more responsive user experience.
- *Reduce Packet Loss*: Emerging technologies such as NFV-SDN optimize network routing, traffic management, and minimize packet loss. This improves reliability, thereby reducing overall transfer time. Furthermore, leveraging and optimizing new protocols such as QUIC can improve network efficiency and provide secure connections with minimal congestion.
- *Bandwidth Optimization*: Caching content closer to end-users at the CDN-edge alleviates congestion on the main network, optimizes network bandwidth and ensures consistent performance even during peak traffic periods.
- *Efficient Adaptation*: The AI-driven adaptation algorithm analyzes application, network, and user behavior. By predicting what clients are likely to request next, it can prefetch content and optimize streaming quality, further improving user satisfaction.
- *Security and Illegal Broadcasting*: Beyond security threats, deploying AI-powered intrusion detection systems and anomaly detection algorithms, along with efficient watermarking technology, help eliminates fraudulent sessions and saves network bandwidth. This ultimately enhances in end-user satisfaction.

Overall, on the server side, content-aware streaming utilizes content characteristics for more efficient encoding and bitrate selection. Machine learning approaches enhance network resource utilization and assist clients achieve smoother adaptation. Reducing latency to near real-time is facilitated by protocols like WebRTC and Low-Latency HLS (LL-HLS). It's important to note that while WebRTC suffers from scalability issues, LL-HLS requires updates on the legacy client side. Ongoing efforts to improve QUIC are aimed at better supporting video streaming applications.

The popularity of 360° video and augmented reality has grown significantly in recent years. However, this study does not delve deeply into this area due to space constraints. For a more in-depth exploration, we recommend consulting the comprehensive surveys in [29] and [30].

REFERENCES

[1] Cisco. (2020). *Cisco Annual Internet Report (2018-2023) White Paper*. Cisco Systems, Inc., San Jose, CA, USA. [Online]. Available: https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html

[2] *Apple's HTTP Live Streaming*. Accessed: Nov. 2024. [Online]. Available: https://developer.apple.com/streaming/

[3] *Microsoft Smooth Streaming*. Accessed: Nov. 2024. [Online]. Available: https://www.iis.net/downloads/microsoft/smooth-streaming

[4] (Nov. 2024). *Dynamic Adaptive Streaming Over HTTP*. [Online]. Available: https://mpeg.chiariglione.org/standards/mpeg-dash

[5] A. Bentaleb, M. Lim, M. N. Akcay, A. C. Begen, S. Hammoudi, and R. Zimmermann, "Toward one-second latency: Evolution of live media streaming," 2023, *arXiv:2310.03256*.

[6] N. Abbas, Y. Zhang, A. Taherkordi, and T. Skeie, "Mobile edge computing: A survey," *IEEE Internet Things J.*, vol. 5, no. 1, pp. 450–465, Feb. 2018.

[7] P. Mach and Z. Becvar, "Mobile edge computing: A survey on architecture and computation offloading," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 3, pp. 1628–1656, 3rd Quart., 2017.

[8] A. Filali, A. Abouaomar, S. Cherkaoui, A. Kobbane, and M. Guizani, "Multi-access edge computing: A survey," *IEEE Access*, vol. 8, pp. 197017–197046, 2020.

[9] R. Mijumbi, J. Serrat, J.-L. Gorricho, N. Bouten, F. De Turck, and R. Boutaba, "Network function virtualization: State-of-the-art and research challenges," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 1, pp. 236–262, 1st Quart., 2016.

[10] B. Yi, X. Wang, K. Li, S. K. Das, and M. Huang, "A comprehensive survey of network function virtualization," *Comput. Netw.*, vol. 133, pp. 212–262, Mar. 2018.

[11] S. Yang, F. Li, S. Trajanovski, R. Yahyapour, and X. Fu, "Recent advances of resource allocation in network function virtualization," *IEEE Trans. Parallel Distrib. Syst.*, vol. 32, no. 2, pp. 295–314, Feb. 2021.

[12] D. Kreutz, F. M. V. Ramos, P. E. Veríssimo, C. E. Rothenberg, S. Azodolmolky, and S. Uhlig, "Software-defined networking: A comprehensive survey," *Proc. IEEE*, vol. 103, no. 1, pp. 14–76, Jan. 2015.

[13] B. A. A. Nunes, M. Mendonca, X.-N. Nguyen, K. Obraczka, and T. Turletti, "A survey of software-defined networking: Past, present, and future of programmable networks," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 3, pp. 1617–1634, 3rd Quart., 2014.

[14] S. H. Haji, S. R. Zeebaree, R. H. Saeed, S. Y. Ameen, H. M. Shukur, N. Omar, M. A. Sadeeq, Z. S. Ageed, I. M. Ibrahim, and H. M. Yasin, "Comparison of software defined networking with traditional networking," *Asian J. Res. Comput. Sci.*, vol. 9, no. 2, pp. 1–18, 2021.

[15] I. H. Sarker, "Machine learning: Algorithms, real-world applications and research directions," *Social Netw. Comput. Sci.*, vol. 2, no. 3, p. 160, May 2021.

[16] B. Mahesh, "Machine learning algorithms—A review," *Int. J. Sci. Res.*, vol. 9, no. 1, pp. 381–386, Jan. 2020.

[17] S. Baraković and L. Skorin-Kapov, "Survey and challenges of QoE management issues in wireless networks," *J. Comput. Netw. Commun.*, vol. 2013, pp. 1–28, Jan. 2013.

[18] M. Seufert, S. Egger, M. Slanina, T. Zinner, T. Hoßfeld, and P. Tran-Gia, "A survey on quality of experience of HTTP adaptive streaming," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 1, pp. 469–492, 1st Quart., 2015.

[19] T. Zhao, Q. Liu, and C. W. Chen, "QoE in video transmission: A user experience-driven strategy," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 1, pp. 285–302, 1st Quart., 2017.

[20] G.-M. Su, X. Su, Y. Bai, M. Wang, A. V. Vasilakos, and H. Wang, "QoE in video streaming over wireless networks: Perspectives and research challenges," *Wireless Netw.*, vol. 22, no. 5, pp. 1571–1593, Jul. 2016.

[21] L. Skorin-Kapov, M. Varela, T. Hoßfeld, and K.-T. Chen, "A survey of emerging concepts and challenges for QoE management of multimedia services," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 14, no. 2s, pp. 1–29, Apr. 2018.

[22] A. Benmir, A. Korichi, A. Bourouis, and M. Alreshoodi, "Survey on QoE/QoS correlation models for video streaming over vehicular ad-hoc networks," *J. Comput. Inf. Technol.*, vol. 26, no. 4, pp. 267–287, 2018.

[23] S. Petrangeli, J. V. D. Hooft, T. Wauters, and F. D. Turck, "Quality of experience-centric management of adaptive video streaming services: Status and challenges," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 14, no. 2s, pp. 1–29, Apr. 2018.

[24] A. Bentaleb, B. Taani, A. C. Begen, C. Timmerer, and R. Zimmermann, "A survey on bitrate adaptation schemes for streaming media over HTTP," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 1, pp. 562–585, 1st Quart., 2019.

[25] A. A. Barakabitze, N. Barman, A. Ahmad, S. Zadtootaghaj, L. Sun, M. G. Martini, and L. Atzori, "QoE management of multimedia streaming services in future networks: A tutorial and survey," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 1, pp. 526–565, 1st Quart., 2020.

[26] N. Barman and M. G. Martini, "QoE modeling for HTTP adaptive video streaming—A survey and open challenges," *IEEE Access*, vol. 7, pp. 30831–30859, 2019.

[27] I. Sousa, M. P. Queluz, and A. Rodrigues, "A survey on QoE-oriented wireless resources scheduling," *J. Netw. Comput. Appl.*, vol. 158, May 2020, Art. no. 102594.

[28] A. A. Barakabitze, A. Ahmad, R. Mijumbi, and A. Hines, "5G network slicing using SDN and NFV: A survey of taxonomy, architectures and future challenges," *Comput. Netw.*, vol. 167, Feb. 2020, Art. no. 106984.

[29] A. Yaqoob, T. Bi, and G.-M. Muntean, "A survey on adaptive 360° video streaming: Solutions, challenges and opportunities," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 4, pp. 2801–2838, 4th Quart., 2020.

[30] Y. Siriwardhana, P. Porambage, M. Liyanage, and M. Ylianttila, "A survey on mobile augmented reality with 5G mobile edge computing: Architectures, applications, and technical aspects," *IEEE Commun. Surveys Tuts.*, vol. 23, no. 2, pp. 1160–1192, 2nd Quart., 2021.

[31] X. Jiang, F. R. Yu, T. Song, and V. C. M. Leung, "A survey on multi-access edge computing applied to video streaming: Some research issues and challenges," *IEEE Commun. Surveys Tuts.*, vol. 23, no. 2, pp. 871–903, 2nd Quart., 2021.

[32] G. Kougioumtzidis, V. Poulkov, Z. D. Zaharis, and P. I. Lazaridis, "A survey on multimedia services QoE assessment and machine learning-based prediction," *IEEE Access*, vol. 10, pp. 19507–19538, 2022.

[33] K. Brunnström, S. A. Beker, K. De Moor, A. Dooms, S. Egger, M.-N. Garcia, T. Hossfeld, S. Jumisko-Pyykkö, C. Keimel, and M.-C. Larabi, "Qualinet white paper on definitions of quality of experience," Lausanne, Switzerland, Version 1.2, Mar. 2013. [Online]. Available: https://www.qualinet.eu/wp-content/uploads/2021/04/QoE_whitepaper_v1.2.pdf

[34] T. Hoßfeld, L. Skorin-Kapov, P. E. Heegaard, and M. Varela, "Definition of QoE fairness in shared systems," *IEEE Commun. Lett.*, vol. 21, no. 1, pp. 184–187, Jan. 2017.

[35] C. Li, M. Lim, A. Bentaleb, and R. Zimmermann, "A real-time blind quality-of-experience assessment metric for HTTP adaptive streaming," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2023, pp. 1661–1666.

[36] M. Jalil Piran, Q.-V. Pham, S. M. R. Islam, S. Cho, B. Bae, D. Y. Suh, and Z. Han, "Multimedia communication over cognitive radio networks from QoS/QoE perspective: A comprehensive survey," *J. Netw. Comput. Appl.*, vol. 172, Dec. 2020, Art. no. 102759.

[37] M. Alreshoodi and J. Woods, "Survey on QoE\QoS correlation models for multimedia services," 2013, *arXiv:1306.0221*.

[38] Z. Duanmu, W. Liu, Z. Li, D. Chen, Z. Wang, Y. Wang, and W. Gao, "Assessing the quality-of-experience of adaptive bitrate video streaming," 2020, *arXiv:2008.08804*.

[39] P. Gao, P. Zhang, and A. Smolic, "Quality assessment for omnidirectional video: A spatio-temporal distortion modeling approach," *IEEE Trans. Multimedia*, vol. 24, pp. 1–16, 2022.

[40] G. J. Sullivan, P. N. Topiwala, and A. Luthra, "The H. 264/AVC advanced video coding standard: Overview and introduction to the fidelity range extensions," *Proc. SPIE*, vol. 5558, pp. 454–474, Nov. 2004.

[41] P. Akyazi and T. Ebrahimi, "Comparison of compression efficiency between HEVC/H.265, VP9 and AV1 based on subjective quality assessments," in *Proc. 10th Int. Conf. Quality Multimedia Exper. (QoMEX)*, May 2018, pp. 1–6.

[42] I. Mansri, N. Doghmane, N. Kouadria, S. Harize, and A. Bekhouch, "Comparative evaluation of VVC, HEVC, H.264, AV1, and VP9 encoders for low-delay video applications," in *Proc. 4th Int. Conf. Multimedia Comput., Netw. Appl. (MCNA)*, Oct. 2020, pp. 38–43.

[43] Y. Chen et al., "An overview of coding tools in AV1: The first video codec from the alliance for open media," *APSIPA Trans. Signal Inf. Process.*, vol. 9, no. 1, p. 6, 2020.

[44] M. Viitanen, J. Sainio, A. Mercat, A. Lemmetti, and J. Vanne, "From HEVC to VVC: The first development steps of a practical intra video encoder," *IEEE Trans. Consum. Electron.*, vol. 68, no. 2, pp. 139–148, May 2022.

[45] I. Unanue, I. Urteaga, R. Husemann, J. Del Ser, V. Roesler, A. Rodríguez, and P. Sánchez, "A tutorial on H. 264/SVC scalable video coding and its tradeoff between quality, coding efficiency and performance," in *Recent Advances on Video Coding*. InTech, 2011, pp. 9–15.

[46] X. Ji, J. Huang, M. Chiang, G. Lafruit, and F. Catthoor, "Scheduling and resource allocation for SVC streaming over OFDM downlink systems," in *Recent Advances on Video Coding*, J. D. S. Lorente, Ed., Rijeka, Croatia: InTechOpen, 2011, ch. 10, doi: 10.5772/21292.

[47] A. Elgabli, V. Aggarwal, S. Hao, F. Qian, and S. Sen, "LBP: Robust rate adaptation algorithm for SVC video streaming," 2018, *arXiv:1805.00041*.

[48] J. Jiang, V. Sekar, and H. Zhang, "Improving fairness, efficiency, and stability in HTTP-based adaptive video streaming with FESTIVE," in *Proc. 8th Int. Conf. Emerg. Netw. Exp. Technol.* New York, NY, USA: ACM, Dec. 2012, pp. 97–108, doi: 10.1145/2413176.2413189.

[49] C.-C. Su, J. J. Yao, and H. H. Chen, "H.264/AVC-based multiple description coding scheme," in *Proc. IEEE Int. Conf. Image Process.*, Jan. 2007, pp. IV-265–IV-268.

[50] M. Nguyen, H. Amirpour, C. Timmerer, and H. Hellwagner, "Scalable high efficiency video coding based HTTP adaptive streaming over QUIC," in *Proc. Workshop Evol., Perform., Interoperability QUIC*, Aug. 2020, pp. 28–34.

[51] S. Xu, S. Sen, and Z. M. Mao, "CSI: Inferring mobile ABR video adaptation behavior under HTTPS and QUIC," in *Proc. 15th Eur. Conf. Comput. Syst.*, Apr. 2020, pp. 1–16.

[52] T. Shreedhar, R. Panda, S. Podanev, and V. Bajpai, "Evaluating QUIC performance over Web, cloud storage, and video workloads," *IEEE Trans. Netw. Service Manage.*, vol. 19, no. 2, pp. 1366–1381, Jun. 2022.

[53] Z. Gurel, T. Erkilic Civelek, and A. C. Begen, "Need for low latency: Media over QUIC," in *Proc. 2nd Mile-High Video Conf.*, May 2023, pp. 139–140.

[54] R. S. Kalan, R. Farahani, E. Karsli, C. Timmerer, and H. Hellwagner, "Towards low latency live streaming: Challenges in a real-world deployment," in *Proc. 13th ACM Multimedia Syst. Conf.*, Jun. 2022, pp. 315–318.

[55] R. Farahani, "Network-assisted delivery of adaptive video streaming services through CDN, SDN, and MEC," 2024, *arXiv:2403.16951*.

[56] R. S. Kalan, M. Sayit, and S. Clayman, "Optimal cache placement and migration for improving the performance of virtualized SAND," in *Proc. IEEE Conf. Netw. Softwarization (NetSoft)*, Jun. 2019, pp. 78–83.

[57] E. Thomas, M. O. van Deventer, M.-L. Champel, O. Oyman, T. Stockhammer, and A. C. Begen, "Applications and deployments of server and network assisted DASH (SAND)," in *Proc. IBC Conf.*, 2016, pp. 1–8.

[58] R. S. Kalan, M. Sayit, and A. C. Begen, "Implementation of SAND architecture using SDN," in *Proc. IEEE Conf. Netw. Function Virtualization Softw. Defined Netw. (NFV-SDN)*, Nov. 2018, pp. 1–6.

[59] (Nov. 2024). *Common Media Application Format*. [Online]. Available: https://mpeg.chiariglione.org/standards/mpeg-a/common-media-application-format

[60] C. Liu, I. Bouazizi, and M. Gabbouj, "Rate adaptation for adaptive HTTP streaming," in *Proc. 2nd Annu. ACM Conf. Multimedia Syst.*, Feb. 2011, pp. 169–174.

[61] C. Liu, I. Bouazizi, M. M. Hannuksela, and M. Gabbouj, "Rate adaptation for dynamic adaptive streaming over HTTP in content distribution network," *Signal Process., Image Commun.*, vol. 27, no. 4, pp. 288–311, Apr. 2012.

[62] Z. Li, X. Zhu, J. Gahm, R. Pan, H. Hu, A. C. Begen, and D. Oran, "Probe and adapt: Rate adaptation for HTTP video streaming at scale," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 4, pp. 719–733, Apr. 2014.

[63] X. Xie, X. Zhang, S. Kumar, and L. E. Li, "PiStream: Physical layer informed adaptive video streaming over LTE," in *Proc. 21st Annu. Int. Conf. Mobile Comput. Netw.*, Sep. 2015, pp. 413–425.

[64] M. Xiao, V. Swaminathan, S. Wei, and S. Chen, "DASH2M: Exploring HTTP/2 for internet streaming to mobile devices," in *Proc. 24th ACM Int. Conf. Multimedia*, Oct. 2016, pp. 22–31.

[65] S. Wei and V. Swaminathan, "Low latency live video streaming over HTTP 2.0," in *Proc. Netw. Operating Syst. Support Digit. Audio Video Workshop*, Mar. 2014, pp. 37–42.

[66] B. Taani and R. Zimmermann, "Spatio-temporal analysis of bandwidth maps for geo-predictive video streaming in mobile environments," in *Proc. 24th ACM Int. Conf. Multimedia*, Oct. 2016, pp. 888–897.

[67] T.-Y. Huang, R. Johari, N. McKeown, M. Trunnell, and M. Watson, "A buffer-based approach to rate adaptation: Evidence from a large video streaming service," in *Proc. ACM Conf. SIGCOMM*, Aug. 2014, pp. 187–198.

[68] C. Sieber, T. Hossfeld, T. Zinner, P. Tran-Gia, and C. Timmerer, "Implementation and user-centric comparison of a novel adaptation logic for DASH with SVC," in *Proc. IFIP/IEEE Int. Symp. Integr. Netw. Manage. (IM)*, May 2013, pp. 1318–1323.

[69] K. Spiteri, R. Urgaonkar, and R. K. Sitaraman, "BOLA: Near-optimal bitrate adaptation for online videos," in *Proc. 35th Annu. IEEE Int. Conf. Comput. Commun.*, Apr. 2016, pp. 1–9.

[70] P. K. Yadav, A. Shafiei, and W. T. Ooi, "QUETRA: A queuing theory approach to DASH rate adaptation," in *Proc. 25th ACM Int. Conf. Multimedia*, Oct. 2017, pp. 1130–1138.

[71] Y. Qin, R. Jin, S. Hao, K. R. Pattipati, F. Qian, S. Sen, C. Yue, and B. Wang, "A control theoretic approach to ABR video streaming: A fresh look at PID-based rate adaptation," *IEEE Trans. Mobile Comput.*, vol. 19, no. 11, pp. 2505–2519, Nov. 2020.

[72] X. Yin, A. Jindal, V. Sekar, and B. Sinopoli, "A control-theoretic approach for dynamic adaptive video streaming over HTTP," in *Proc. ACM Conf. Special Interest Group Data Commun.*, Aug. 2015, pp. 325–338.

[73] P. Juluri, V. Tamarapalli, and D. Medhi, "SARA: Segment aware rate adaptation algorithm for dynamic adaptive streaming over HTTP," in *Proc. IEEE Int. Conf. Commun. Workshop (ICCW)*, Jun. 2015, pp. 1765–1770.

[74] C. Wang, A. Rizk, and M. Zink, "SQUAD: A spectrum-based quality adaptation for dynamic adaptive streaming over HTTP," in *Proc. 7th Int. Conf. Multimedia Syst.*, May 2016, pp. 1–12.

[75] A. Beben, P. Wiśniewski, J. M. Batalla, and P. Krawiec, "ABMA+: Lightweight and efficient algorithm for HTTP adaptive streaming," in *Proc. 7th Int. Conf. Multimedia Syst.*, May 2016, pp. 1–11.

[76] L. De Cicco, V. Caldaralo, V. Palmisano, and S. Mascolo, "ELASTIC: A client-side controller for dynamic adaptive streaming over HTTP (DASH)," in *Proc. 20th Int. Packet Video Workshop*, Dec. 2013, pp. 1–8.

[77] J. Jiang, V. Sekar, and H. Zhang, "Improving fairness, efficiency, and stability in HTTP-based adaptive video streaming with festive," *IEEE/ACM Trans. Netw.*, vol. 22, no. 1, pp. 326–340, Feb. 2014.

[78] H. Mao, R. Netravali, and M. Alizadeh, "Neural adaptive video streaming with pensieve," in *Proc. Conf. ACM Special Interest Group Data Commun.*, Aug. 2017, pp. 197–210.

[79] C. Zhou, C.-W. Lin, and Z. Guo, "MDASH: A Markov decision-based rate adaptation approach for dynamic HTTP streaming," *IEEE Trans. Multimedia*, vol. 18, no. 4, pp. 738–751, Apr. 2016.

[80] M. Gadaleta, F. Chiariotti, M. Rossi, and A. Zanella, "D-DASH: A deep Q-learning framework for DASH video streaming," *IEEE Trans. Cognit. Commun. Netw.*, vol. 3, no. 4, pp. 703–718, Dec. 2017.

[81] R.-X. Zhang, T. Huang, M. Ma, H. Pang, X. Yao, C. Wu, and L. Sun, "Enhancing the crowdsourced live streaming: A deep reinforcement learning approach," in *Proc. 29th ACM Workshop Netw. Operating Syst. Support Digit. Audio Video*, Jun. 2019, pp. 55–60.

[82] L. De Cicco, S. Mascolo, and V. Palmisano, "Feedback control for adaptive live video streaming," in *Proc. 2nd Annu. ACM Conf. Multimedia Syst.*, Feb. 2011, pp. 145–156.

[83] L. Zhang, S. Wang, and R. N. Chang, "QCSS: A QoE-aware control plane for adaptive streaming service over mobile edge computing infrastructures," in *Proc. IEEE Int. Conf. Web Services (ICWS)*, Jul. 2018, pp. 139–146.

[84] J. Bruneau-Queyreix, M. Lacaud, D. Negru, J. M. Batalla, and E. Borcoci, "MS-stream: A multiple-source adaptive streaming solution enhancing consumer's perceived quality," in *Proc. 14th IEEE Annu. Consum. Commun. Netw. Conf. (CCNC)*, Jan. 2017, pp. 427–434.

[85] J. Chen, R. Mahindra, M. A. Khojastepour, S. Rangarajan, and M. Chiang, "A scheduling framework for adaptive video delivery over cellular networks," in *Proc. 19th Annu. Int. Conf. Mobile Comput. Netw.*, 2013, p. 389.

[86] V. Joseph and G. de Veciana, "NOVA: QoE-driven optimization of DASH-based video delivery in networks," in *Proc. IEEE IEEE Conf. Comput. Commun.*, Apr. 2014, pp. 82–90.

[87] T. Huang, R.-X. Zhang, C. Zhou, and L. Sun, "QARC: Video quality aware rate control for real-time video streaming based on deep reinforcement learning," in *Proc. 26th ACM Int. Conf. Multimedia*, Oct. 2018, pp. 1208–1216.

[88] S. Petrangeli, J. Famaey, M. Claeys, S. Latré, and F. De Turck, "QoE-driven rate adaptation heuristic for fair adaptive video streaming," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 12, no. 2, pp. 1–24, Mar. 2016.

[89] R. K. P. Mok, X. Luo, E. W. W. Chan, and R. K. C. Chang, "QDASH: A QoE-aware DASH system," in *Proc. 3rd Multimedia Syst. Conf.*, Feb. 2012, pp. 11–22.

[90] V. Krishnamoorthi, N. Carlsson, E. Halepovic, and E. Petajan, "BUFFEST: Predicting buffer conditions and real-time requirements of HTTP(S) adaptive streaming clients," in *Proc. 8th ACM Multimedia Syst. Conf.*, Jun. 2017, pp. 76–87.

[91] B. Han, F. Qian, L. Ji, and V. Gopalakrishnan, "MP-DASH: Adaptive video streaming over preference-aware multipath," in *Proc. 12th Int. Conf. Emerg. Netw. Exp. Technol.* New York, NY, USA: ACM, Dec. 2016, pp. 129–143.

[92] K. Du, A. Pervaiz, X. Yuan, A. Chowdhery, Q. Zhang, H. Hoffmann, and J. Jiang, "Server-driven video streaming for deep learning inference," in *Proc. Annu. Conf. ACM Special Interest Group Data Commun. Appl., Technol., Architectures, Protocols Comput. Commun.*, Jul. 2020, pp. 557–570.

[93] F. Tashtarian, A. Bentaleb, H. Amirpour, B. Taraghi, C. Timmerer, H. Hellwagner, and R. Zimmermann, "LALISA: Adaptive bitrate ladder optimization in HTTP-based adaptive live streaming," in *Proc. NOMS-IEEE/IFIP Netw. Oper. Manage. Symp.*, May 2023, pp. 1–9.

[94] F. Tashtarian, A. Bentaleb, H. Amirpour, S. Gorinsky, J. Jiang, H. Hellwagner, and C. Timmerer, "ARTEMIS: Adaptive bitrate ladder optimization for live video streaming," in *Proc. 21st USENIX Symp. Networked Syst. Design Implement. (NSDI)*, 2024, pp. 591–611.

[95] H. Amirpour, E. Çetinkaya, C. Timmerer, and M. Ghanbari, "Towards optimal multirate encoding for http adaptive streaming," in *Proc. 27th Int. Conf. MultiMedia Modeling*, Prague, Czech Republic. Cham, Switzerland: Springer, 2021, pp. 469–480.

[96] E. Çetinkaya, H. Amirpour, C. Timmerer, and M. Ghanbari, "FaME-ML: Fast multirate encoding for HTTP adaptive streaming using machine learning," in *Proc. IEEE Int. Conf. Vis. Commun. Image Process. (VCIP)*, Dec. 2020, pp. 87–90.

[97] E. Cetinkaya, H. Amirpour, C. Timmerer, and M. Ghanbari, "C," *IEEE Open J. Signal Process.*, vol. 2, pp. 484–495, 2021.

[98] O. S. Faragallah, W. El-Shafai, A. I. Sallam, I. Elashry, E.-S.-M. El-Rabaie, A. Afifi, M. A. AlZain, J. F. Al-Amri, F. E. A. El-Samie, and H. S. El-sayed, "Cybersecurity framework of hybrid watermarking and selective encryption for secure HEVC communication," *J. Ambient Intell. Humanized Comput.*, vol. 13, no. 2, pp. 1215–1239, Feb. 2022.

[99] (Nov. 2024). *DASH-IF IOP Watermarking*. [Online]. Available: https://dashif.org/docs /DASH-IF-IOP_OTT-Watermarking.pdf

[100] Q. Chang, L. Huang, S. Liu, H. Liu, T. Yang, and Y. Wang, "Blind robust video watermarking based on adaptive region selection and channel reference," in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 2344–2350.

[101] R. S. Kalan and E. Karslı, "Illegal broadcasting: A way for protecting revenue," in *Proc. 14th Int. Conf. Netw. Future (NoF)*, Oct. 2023, pp. 57–61.

[102] S. Gaj, A. Sur, and P. K. Bora, "Prediction mode based H.265/HEVC video watermarking resisting re-compression attack," *Multimedia Tools Appl.*, vol. 79, nos. 25–26, pp. 18089–18119, Jul. 2020.

[103] H. Ding, R. Tao, J. Sun, J. Liu, F. Zhang, X. Jiang, and J. Li, "A compressed-domain robust video watermarking against recompression attack," *IEEE Access*, vol. 9, pp. 35324–35337, 2021.

[104] H. Mareen, G. Van Wallendael, and P. Lambert, "Implementation-free forensic watermarking for adaptive streaming with A/B watermarking," in *Proc. 6th Int. Congr. Inf. Commun. Technol.*, vol. 1, London, U.K. Cham, Switzerland: Springer, 2022, pp. 325–339.

[105] O. S. Faragallah, A. Afifi, H. S. El-Sayed, M. A. Alzain, J. F. Al-Amri, F. E. A. El-Samie, and W. El-Shafai, "Efficient HEVC integrity verification scheme for multimedia cybersecurity applications," *IEEE Access*, vol. 8, pp. 167069–167089, 2020.

[106] L. Tian, H. Dai, and C. Li, "A semi-fragile video watermarking algorithm based on chromatic residual DCT," *Multimedia Tools Appl.*, vol. 79, nos. 3–4, pp. 1759–1779, Jan. 2020.

[107] X.-C. Sun, Z.-M. Lu, Z. Wang, and Y.-L. Liu, "A geometrically robust multi-bit video watermarking algorithm based on 2-D DFT," *Multimedia Tools Appl.*, vol. 80, no. 9, pp. 13491–13511, Apr. 2021.

[108] D. Rodríguez Galiano, A. A. Del Barrio, G. Botella, and D. Cuesta, "Efficient embedding and retrieval of information for high-resolution videos coded with HEVC," *Comput. Electr. Eng.*, vol. 81, Jan. 2020, Art. no. 106541.

[109] X. Wu, P. Ma, Z. Jin, Y. Wu, W. Han, and W. Ou, "A novel zero-watermarking scheme based on NSCT-SVD and blockchain for video copyright," *EURASIP J. Wireless Commun. Netw.*, vol. 2022, no. 1, p. 20, Dec. 2022.

[110] P. R. Padma, T. Aditya, P. Ayush, and K. Siddharth, "Identifying video tampering using watermarked blockchain," *Int. J. Performability Eng.*, vol. 17, no. 8, p. 722, 2021.

[111] A. Mishra, M. Bansal, and A. Sharma, "Video watermarking of live streamed MPEG-4 frames using ELM-fuzzy-PSO hybrid scheme," *Multimedia Tools Appl.*, vol. 83, no. 14, pp. 41997–42035, Oct. 2023.
[112] X. Luo, Y. Li, H. Chang, C. Liu, P. Milanfar, and F. Yang, "DVMark: A deep multiscale framework for video watermarking," *IEEE Trans. Image Process.*, early access, Mar. 28, 2023, doi: 10.1109/TIP.2023.3251737.
[113] A. A. Elrowayati, M. A. Alrshah, M. F. L. Abdullah, and R. Latip, "HEVC watermarking techniques for authentication and copyright applications: Challenges and opportunities," *IEEE Access*, vol. 8, pp. 114172–114189, 2020.
[114] A. Lekharu, K. Y. Moulii, A. Sur, and A. Sarkar, "Deep learning based prediction model for adaptive video streaming," in *Proc. Int. Conf. Commun. Syst. Netw. (COMSNETS)*, Jan. 2020, pp. 152–159.
[115] Y. Zhang, Y. Zhang, Y. Wu, Y. Tao, K. Bian, P. Zhou, L. Song, and H. Tuo, "Improving quality of experience by adaptive video streaming with super-resolution," in *Proc. IEEE Conf. Comput. Commun.*, Jul. 2020, pp. 1957–1966.
[116] X. Wei, M. Zhou, S. Kwong, H. Yuan, S. Wang, G. Zhu, and J. Cao, "Reinforcement learning-based QoE-oriented dynamic adaptive streaming framework," *Inf. Sci.*, vol. 569, pp. 786–803, Aug. 2021.
[117] H. Wang, K. Wu, J. Wang, and G. Tang, "Rldish: Edge-assisted QoE optimization of HTTP live streaming with reinforcement learning," in *Proc. IEEE Conf. Comput. Commun.*, Jul. 2020, pp. 706–715.
[118] X. Ma, Q. Li, Y. Jiang, G.-M. Muntean, and L. Zou, "Learning-based joint QoE optimization for adaptive video streaming based on smart edge," *IEEE Trans. Netw. Service Manage.*, vol. 19, no. 2, pp. 1789–1806, Jun. 2022.
[119] H. Zhang, L. Dong, G. Gao, H. Hu, Y. Wen, and K. Guan, "DeepQoE: A multimodal learning framework for video quality of experience (QoE) prediction," *IEEE Trans. Multimedia*, vol. 22, no. 12, pp. 3210–3223, Dec. 2020.
[120] S. Altamimi and S. Shirmohammadi, "QoE-fair DASH video streaming using server-side reinforcement learning," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 16, no. 2s, pp. 1–21, Apr. 2020.
[121] F. Tashtarian, R. Falanji, A. Bentaleb, A. Erfanian, P. S. Mashhadi, C. Timmerer, H. Hellwagner, and R. Zimmermann, "Quality optimization of live streaming services over HTTP with reinforcement learning," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2021, pp. 1–6.
[122] W. Zhou, X. Min, H. Li, and Q. Jiang, "A brief survey on adaptive video streaming quality assessment," *J. Vis. Commun. Image Represent.*, vol. 86, Jul. 2022, Art. no. 103526.
[123] M. Naresh, V. Das, P. Saxena, and M. Gupta, "Deep reinforcement learning based QoE-aware actor-learner architectures for video streaming in IoT environments," *Computing*, vol. 104, no. 7, pp. 1527–1550, Jul. 2022.
[124] Y. Wang, M. Agarwal, T. Lan, and V. Aggarwal, "Learning-based online QoE optimization in multi-agent video streaming," *Algorithms*, vol. 15, no. 7, p. 227, Jun. 2022.
[125] F. Y. Yan, H. Ayers, C. Zhu, S. Fouladi, J. Hong, K. Zhang, P. Levis, and K. Winstein, "Learning in situ: A randomized experiment in video streaming," in *Proc. 17th USENIX Symp. Networked Syst. Design Implement. (NSDI)*, 2020, pp. 495–511.
[126] X. Zuo, J. Yang, M. Wang, and Y. Cui, "Adaptive bitrate with user-level QoE preference for video streaming," in *Proc. IEEE Conf. Comput. Commun.*, May 2022, pp. 1279–1288.
[127] C. Qiao, J. Wang, and Y. Liu, "Beyond QoE: Diversity adaptation in video streaming at the edge," *IEEE/ACM Trans. Netw.*, vol. 29, no. 1, pp. 289–302, Feb. 2021.
[128] R. S. Kalan, "Enhancing quality of experience by monitoring end-user streaming behavior," in *Proc. IEEE Int. Conf. Dependable, Autonomic Secure Comput., Int. Conf. Pervasive Intell. Comput., Int. Conf. Cloud Big Data Comput., Int. Conf. Cyber Sci. Technol. Congr. (DASC/PiCom/CBDCom/CyberSciTech)*, Nov. 2023, pp. 511–514.
[129] F. Laiche, A. Ben Letaifa, I. Elloumi, and T. Aguili, "When machine learning algorithms meet user engagement parameters to predict video QoE," *Wireless Pers. Commun.*, vol. 116, no. 3, pp. 2723–2741, Feb. 2021.
[130] F. Loh, F. Poignée, F. Wamser, F. Leidinger, and T. Hoßfeld, "Uplink vs. downlink: Machine learning-based quality prediction for HTTP adaptive video streaming," *Sensors*, vol. 21, no. 12, p. 4172, Jun. 2021.
[131] N. Souane, M. Bourenane, and Y. Douga, "Deep reinforcement learning-based approach for video streaming: Dynamic adaptive video streaming over HTTP," *Appl. Sci.*, vol. 13, no. 21, p. 11697, Oct. 2023.
[132] K. Khan, "A framework for meta-learning in dynamic adaptive streaming over http," *Int. J. Comput.*, vol. 12, no. 2, pp. 1–10, 2023.
[133] W. Li, J. Huang, S. Wang, C. Wu, S. Liu, and J. Wang, "An apprenticeship learning approach for adaptive video streaming based on chunk quality and user preference," *IEEE Trans. Multimedia*, vol. 25, pp. 2488–2502, 2022.
[134] H. Mao, S. Chen, D. Dimmery, S. Singh, D. Blaisdell, Y. Tian, M. Alizadeh, and E. Bakshy, "Real-world video adaptation with reinforcement learning," 2020, *arXiv:2008.12858*.
[135] C. Müller, S. Steglich, S. Groß, and P. Kremer, "Machine-learning based VMAF prediction for HDR video content," in *Proc. 14th ACM Multimedia Syst. Conf.*, Jun. 2023, pp. 328–332.
[136] A. Bentaleb, M. Lim, M. N. Akcay, A. C. Begen, and R. Zimmermann, "Meta reinforcement learning for rate adaptation," in *Proc. IEEE Conf. Comput. Commun.*, May 2023, pp. 1–10.
[137] A. Bentaleb, M. Lim, M. N. Akcay, A. C. Begen, and R. Zimmermann, "Bitrate adaptation and guidance with meta reinforcement learning," *IEEE Trans. Mobile Comput.*, vol. 23, no. 11, pp. 10378–10392, Nov. 2024.
[138] V. V. Menon, R. Farahani, P. T. Rajendran, S. Afzal, K. Schoeffmann, and C. Timmerer, "Energy-efficient multi-codec bitrate-ladder estimation for adaptive video streaming," in *Proc. IEEE Int. Conf. Vis. Commun. Image Process. (VCIP)*, Dec. 2023, pp. 1–5.
[139] X. Chen, T. Tan, G. Cao, and T. F. L. Porta, "Context-aware and energy-aware video streaming on smartphones," *IEEE Trans. Mobile Comput.*, vol. 21, no. 3, pp. 862–877, Mar. 2022.
[140] B. O. Turkkan, T. Dai, A. Raman, T. Kosar, C. Chen, M. F. Bulut, J. Zola, and D. Sow, "GreenABR: Energy-aware adaptive bitrate streaming with deep reinforcement learning," in *Proc. 13th ACM Multimedia Syst. Conf.*, Jun. 2022, pp. 150–163.
[141] A. Mondal, B. Palit, S. Khandelia, N. Pal, J. Jayatheerthan, K. Paul, N. Ganguly, and S. Chakraborty, "EnDASH–A mobility adapted energy efficient ABR video streaming for cellular networks," in *Proc. IFIP Netw. Conf. (Networking)*, Jun. 2020, pp. 127–135.
[142] B. Palit, A. Sen, A. Mondal, A. Zunaid, J. Jayatheerthan, and S. Chakraborty, "Improving UE energy efficiency through network-aware video streaming over 5G," *IEEE Trans. Netw. Service Manage.*, vol. 20, no. 3, pp. 3487–3500, Sep. 2023.
[143] Y. Guan, Y. Zhang, B. Wang, K. Bian, X. Xiong, and L. Song, "PERM: Neural adaptive video streaming with multi-path transmission," in *Proc. IEEE Conf. Comput. Commun.*, Jul. 2020, pp. 1103–1112.
[144] L. Zhong, X. Ji, Z. Wang, J. Qin, and G.-M. Muntean, "A Q-learning driven energy-aware multipath transmission solution for 5G media services," *IEEE Trans. Broadcast.*, vol. 68, no. 2, pp. 559–571, Jun. 2022.
[145] K. Wheatman, F. Mehmeti, M. Mahon, T. La Porta, and G. Cao, "Multi-user competitive energy-aware and QoE-aware video streaming on mobile devices," in *Proc. 16th ACM Symp. QoS Secur. Wireless Mobile Netw.*, Nov. 2020, pp. 47–55.
[146] J. Xie, "Deep Q-learning aided energy-efficient caching and transmission for adaptive bitrate video streaming over dynamic cellular networks," *IEEE Access*, vol. 12, pp. 24232–24242, 2024.
[147] X. Hu, A. Ghosh, X. Liu, Z.-L. Zhang, and N. Shroff, "COREL: Constrained reinforcement learning for video streaming ABR algorithm design over mmWave 5G," in *Proc. IEEE Int. Workshop Tech. Committee Commun. Quality Rel. (CQR)*, Oct. 2023, pp. 1–6.
[148] K. Park, M. Kim, and L. Park, "NeuSaver: Neural adaptive power consumption optimization for mobile video streaming," *IEEE Trans. Mobile Comput.*, vol. 22, no. 11, pp. 6633–6646, Nov. 2022.
[149] S. Wang, S. Bi, and Y. A. Zhang, "Deep reinforcement learning with communication transformer for adaptive live streaming in wireless edge networks," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 1, pp. 308–322, Jan. 2022.
[150] C. Qian, D. Liu, and H. Jiang, "Harmonizing energy efficiency and QoE for brightness scaling-based mobile video streaming," in *Proc. IEEE/ACM 30th Int. Symp. Quality Service (IWQoS)*, Jun. 2022, pp. 1–10.
[151] D. Lorenzi, M. Nguyen, F. Tashtarian, and C. Timmerer, "E-WISH: An energy-aware ABR algorithm for green HTTP adaptive video streaming," in *Proc. 3rd Mile-High Video Conf.*, Feb. 2024, pp. 28–33.

[152] J. Meng, Q. Xu, and Y. C. Hu, "Proactive energy-aware adaptive video streaming on mobile devices," in *Proc. USENIX Annu. Tech. Conf.*, 2021, pp. 303–316.

[153] K. B. Ajeyprasaath and P. Vetrivelan, "Machine learning based classifiers for QoE prediction framework in video streaming over 5G wireless networks," *Comput., Mater. Continua*, vol. 75, no. 1, pp. 1919–1939, 2023.

[154] H. Cai, W. Jing, X. Wen, Z. Lu, and Z. Wang, "MEC-based QoE optimization for adaptive video streaming via satellite backhaul," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, Jun. 2021, pp. 1–7.

[155] M. Taha, A. Canovas, J. Lloret, and A. Ali, "A QoE adaptive management system for high definition video streaming over wireless networks," *Telecommun. Syst.*, vol. 77, no. 1, pp. 63–81, May 2021.

[156] A. Xiao, X. Huang, S. Wu, H. Chen, and L. Ma, "Traffic-aware rate adaptation for improving time-varying QoE factors in mobile video streaming," *IEEE Trans. Netw. Sci. Eng.*, vol. 7, no. 4, pp. 2392–2405, Oct. 2020.

[157] R. Ul Mustafa, S. Ferlin, C. Esteve Rothenberg, D. Raca, and J. J. Quinlan, "A supervised machine learning approach for DASH video QoE prediction in 5G networks," in *Proc. 16th ACM Symp. QoS Secur. Wireless Mobile Netw.*, Nov. 2020, pp. 57–64.

[158] Y. Sani, D. Raca, J. J. Quinlan, and C. J. Sreenan, "SMASH: A supervised machine learning approach to adaptive video streaming over HTTP," in *Proc. 12th Int. Conf. Quality Multimedia Exper. (QoMEX)*, May 2020, pp. 1–6.

[159] W. U. Rahman, M. B. Amin, M. D. Hossain, C. S. Hong, and E.-N. Huh, "QoE optimization for HTTP adaptive streaming: Performance evaluation of MEC-assisted and client-based methods," *J. Vis. Commun. Image Represent.*, vol. 82, Jan. 2022, Art. no. 103415.

[160] N. A. Hafez, M. S. Hassan, and T. Landolsi, "Reinforcement learning-based rate adaptation in dynamic video streaming," *Telecommun. Syst.*, vol. 83, no. 4, pp. 395–407, Aug. 2023.

[161] T. Abar, A. B. Letaifa, and S. E. Asmi, "User behavior-ensemble learning based improving QoE fairness in HTTP adaptive streaming over SDN approach," in *Advances in Computers*, vol. 123. Amsterdam, The Netherlands: Elsevier, 2021, pp. 245–269.

[162] L. Zhong, M. Wang, C. Xu, S. Yang, and G.-M. Muntean, "Decentralized optimization for multicast adaptive video streaming in edge cache-assisted networks," *IEEE Trans. Broadcast.*, vol. 69, no. 3, pp. 812–822, Sep. 2023.

[163] H. Noh, H. Lee, Y. Go, H. Park, J. Lee, J. Kim, and H. Song, "Congestion-aware HTTP adaptive streaming system over SDN-enabled Wi-Fi network," *Concurrency Comput., Pract. Exper.*, vol. 32, no. 21, p. e5310, Nov. 2020.

[164] J. Aguilar-Armijo, E. Çetinkaya, C. Timmerer, and H. Hellwagner, "ECAS-ML: Edge computing assisted adaptation scheme with machine learning for HTTP adaptive streaming," in *Proc. Int. Conf. Multimedia Modelling*. Cham, Switzerland: Springer, 2022, pp. 394–406.

[165] M. Taha, A. Ali, J. Lloret, P. R. L. Gondim, and A. Canovas, "An automated model for the assessment of QoE of adaptive video streaming over wireless networks," *Multimedia Tools Appl.*, vol. 80, no. 17, pp. 26833–26854, Jul. 2021.

[166] X. Ma, Q. Li, L. Zou, J. Peng, J. Zhou, J. Chai, Y. Jiang, and G.-M. Muntean, "QAVA: QoE-aware adaptive video bitrate aggregation for HTTP live streaming based on smart edge computing," *IEEE Trans. Broadcast.*, vol. 68, no. 3, pp. 661–676, Sep. 2022.

[167] M. Kim and K. Chung, "Reinforcement learning-based adaptive streaming scheme with edge computing assistance," *Sensors*, vol. 22, no. 6, p. 2171, Mar. 2022.

[168] R. Bhattacharyya, A. Bura, D. Rengarajan, M. Rumuly, B. Xia, S. Shakkottai, D. Kalathil, R. K. P. Mok, and A. Dhamdhere, "QFlow: A learning approach to high QoE video streaming at the wireless edge," *IEEE/ACM Trans. Netw.*, vol. 30, no. 1, pp. 32–46, Feb. 2022.

[169] M. Kim and K. Chung, "HTTP adaptive streaming scheme based on reinforcement learning with edge computing assistance," *J. Netw. Comput. Appl.*, vol. 213, Apr. 2023, Art. no. 103604.

[170] T. V. Huu, S. Van Pham, T. N. T. Huong, and H.-C. Le, "QoE aware video streaming scheme utilizing GRU-based bandwidth prediction and adaptive bitrate selection for heterogeneous mobile networks," *IEEE Access*, vol. 12, pp. 45785–45795, 2024.

[171] W. Liu, H. Zhang, H. Ding, and D. Yuan, "Delay and energy minimization for adaptive video streaming: A joint edge caching, computing and power allocation approach," *IEEE Trans. Veh. Technol.*, vol. 71, no. 9, pp. 9602–9612, Sep. 2022.

[172] A. Zhang, Q. Li, Y. Chen, X. Ma, L. Zou, Y. Jiang, Z. Xu, and G.-M. Muntean, "Video super-resolution and caching—An edge-assisted adaptive video streaming solution," *IEEE Trans. Broadcast.*, vol. 67, no. 4, pp. 799–812, Dec. 2021.

[173] R. S. Kalan, "Big data: Real-time video streaming and log analytic for improving quality of experience," *Int. J. Future Comput. Commun.*, vol. 13, no. 2, pp. 1–10, 2024.

[174] A. O. Al-Abbasi, V. Aggarwal, and M.-R. Ra, "Multi-tier caching analysis in CDN-based over-the-top video streaming systems," *IEEE/ACM Trans. Netw.*, vol. 27, no. 2, pp. 835–847, Apr. 2019.

[175] F. Wang, F. Wang, J. Liu, R. Shea, and L. Sun, "Intelligent video caching at network edge: A multi-agent deep reinforcement learning approach," in *Proc. IEEE Conf. Comput. Commun.*, Jul. 2020, pp. 2499–2508.

[176] R. Viola, A. Martin, J. Morgade, S. Masneri, M. Zorrilla, P. Angueira, and J. Montalbán, "Predictive CDN selection for video delivery based on LSTM network performance forecasts and cost-effective trade-offs," *IEEE Trans. Broadcast.*, vol. 67, no. 1, pp. 145–158, Mar. 2021.

[177] K. Akpinar and K. A. Hua, "PPNet: Privacy protected CDN-ISP collaboration for QoS-aware multi-CDN adaptive video streaming," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 16, no. 2, pp. 1–23, May 2020.

[178] R. Farahani, A. Bentaleb, E. Çetinkaya, C. Timmerer, R. Zimmermann, and H. Hellwagner, "Hybrid P2P-CDN architecture for live video streaming: An online learning approach," in *Proc. IEEE Global Commun. Conf.*, Dec. 2022, pp. 1911–1917.

[179] S. Budhkar and V. Tamarapalli, "An overlay management strategy to improve QoS in CDN-P2P live streaming systems," *Peer-Peer Netw. Appl.*, vol. 13, no. 1, pp. 190–206, Jan. 2020.

[180] S. Nacakli and A. M. Tekalp, "Controlling P2P-CDN live streaming services at SDN-enabled multi-access edge datacenters," *IEEE Trans. Multimedia*, vol. 23, pp. 3805–3816, 2021.

[181] W. Yang, J. Cao, and F. Wu, "Adaptive video streaming with scalable video coding using multipath QUIC," in *Proc. IEEE Int. Perform., Comput., Commun. Conf. (IPCCC)*, Oct. 2021, pp. 1–7.

[182] A. Bentaleb and S. Harous, "Inferring quality of experience for adaptive video streaming over HTTPS and QUIC," in *Proc. Int. Wireless Commun. Mobile Comput. (IWCMC)*, Jun. 2020, pp. 81–87.

[183] A. Bujari, C. E. Palazzi, G. Quadrio, and D. Ronzani, "Emerging interactive applications over QUIC," in *Proc. IEEE 17th Annu. Consum. Commun. Netw. Conf. (CCNC)*, Jan. 2020, pp. 1–4.

[184] S. Lee and J. Yoo, "Reinforcement learning based multipath QUIC scheduler for multimedia streaming," *Sensors*, vol. 22, no. 17, p. 6333, Aug. 2022.

[185] V. A. Vu and B. Walker, "On the latency of multipath-QUIC in real-time applications," in *Proc. 16th Int. Conf. Wireless Mobile Comput., Netw. Commun. (WiMob)*, Oct. 2020, pp. 1–7.

[186] X. Zhang, S. Jin, Y. He, A. Hassan, Z. M. Mao, F. Qian, and Z.-L. Zhang, "QUIC is not quick enough over fast Internet," in *Proc. ACM Web Conf.*, May 2024, pp. 2713–2722.

[187] A. Mondal and S. Chakraborty, "Does QUIC suit well with modern adaptive bitrate streaming techniques?" *IEEE Netw. Lett.*, vol. 2, no. 2, pp. 85–89, Jun. 2020.

[188] S. Arisu, E. Yildiz, and A. C. Begen, "Game of protocols: Is QUIC ready for prime time streaming?" *Int. J. Netw. Manage.*, vol. 30, no. 3, p. e2063, May 2020.

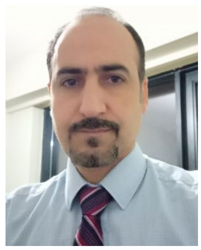
[189] H. K. Ravuri, M. T. Vega, J. D. Van Hooft, T. Wauters, and F. De Turck, "Adaptive partially reliable delivery of immersive media over QUIC-HTTP/3," *IEEE Access*, vol. 11, pp. 38094–38111, 2023.

[190] M. Zverev, P. Garrido, F. Fernández, J. Bilbao, Ö. Alay, S. Ferlin, A. Brunstrom, and R. Agüero, "Robust QUIC: Integrating practical coding in a low latency transport protocol," *IEEE Access*, vol. 9, pp. 138225–138244, 2021.

[191] D. Kreutz, F. Ramos, P. Verissimo, C. E. Rothenberg, S. Azodolmolky, and S. Uhlig, "Software-defined networking: A comprehensive survey," 2014, *arXiv:1406.0440*.

[192] A. Gember-Jacobson, R. Viswanathan, C. Prakash, R. Grandl, J. Khalid, S. Das, and A. Akella, "OpenNF: Enabling innovation in network function control," in *Proc. ACM Conf. SIGCOMM*. New York, NY, USA: ACM, Aug. 2014, pp. 163–174.

[193] X. Xiao, A. Hannan, B. Bailey, and L. M. Ni, "Traffic engineering with MPLS in the internet," *IEEE Netw.*, vol. 14, no. 2, pp. 28–33, Mar. 2000.
[194] R. Jain and S. Paul, "Network virtualization and software defined networking for cloud computing: A survey," *IEEE Commun. Mag.*, vol. 51, no. 11, pp. 24–31, Nov. 2013.
[195] Y. Li and M. Chen, "Software-defined network function virtualization: A survey," *IEEE Access*, vol. 3, pp. 2542–2553, 2015.
[196] B. Han, V. Gopalakrishnan, L. Ji, and S. Lee, "Network function virtualization: Challenges and opportunities for innovations," *IEEE Commun. Mag.*, vol. 53, no. 2, pp. 90–97, Feb. 2015.
[197] J. Chen, J. Chen, and H. Zhang, "DRL-QOR: Deep reinforcement learning-based QoS/QoE-aware adaptive online orchestration in NFV-enabled networks," *IEEE Trans. Netw. Service Manage.*, vol. 18, no. 2, pp. 1758–1774, Jun. 2021.
[198] F. Laiche, A. Ben Letaifa, and T. Aguili, "QoE-aware traffic monitoring based on user behavior in video streaming services," *Concurrency Comput., Pract. Exper.*, vol. 35, no. 11, p. e6678, May 2023.
[199] P. Zhou, Y. Xie, B. Niu, L. Pu, Z. Xu, H. Jiang, and H. Huang, "QoE-aware 3D video streaming via deep reinforcement learning in software defined networking enabled mobile edge computing," *IEEE Trans. Netw. Sci. Eng.*, vol. 8, no. 1, pp. 419–433, Jan. 2021.
[200] A. Erfanian, F. Tashtarian, A. Zabrovskiy, C. Timmerer, and H. Hellwagner, "OSCAR: On optimizing resource utilization in live video streaming," *IEEE Trans. Netw. Service Manage.*, vol. 18, no. 1, pp. 552–569, Mar. 2021.
[201] R. S. Kalan, M. Sayit, and S. Clayman, "Multimedia service management with virtualized cache migration," in *Proc. 23rd Conf. Innov. Clouds, Internet Netw. Workshops (ICIN)*, Feb. 2020, pp. 171–177.
[202] R. S. Kalan, S. Clayman, and M. Sayıt, "VDANE: Using virtualization for improving video quality with server and network assisted DASH," *Int. J. Netw. Manage.*, vol. 32, no. 5, p. e2209, Sep. 2022.
[203] R. S. Kalan, "QoE-driven multimedia streaming leveraging distributed network function virtualization at edge," in *Proc. 8th Int. Conf. Inf. Netw. Technol. (ICINT)*, May 2023, pp. 27–32.
[204] S. Bayhan, S. Maghsudi, and A. Zubow, "EdgeDASH: Exploiting network-assisted adaptive video streaming for edge caching," *IEEE Trans. Netw. Service Manage.*, vol. 18, no. 2, pp. 1732–1745, Jun. 2021.
[205] H. Noh, G. S. Park, Y. Go, S.-H. Shin, Y. Jang, and H. Song, "Energy-efficient HTTP adaptive streaming system over SDN-enabled Wi-Fi APs," *IEEE Access*, vol. 11, pp. 105427–105439, 2023.
[206] A. Hegde, M. Vijayalakshmi, and G. N. Jayalaxmi, "QoE aware video adaptation for video streaming in 5G networks," in *Proc. IEEE Int. Conf. Data Sci. Inf. Syst. (ICDSIS)*, Jul. 2022, pp. 1–8.
[207] R. Farahani, A. Bentaleb, C. Timmerer, M. Shojafar, R. Prodan, and H. Hellwagner, "SARENA: SFC-enabled architecture for adaptive video streaming applications," in *Proc. IEEE Int. Conf. Commun.*, May 2023, pp. 864–870.
[208] D. E. Henni, A. Ghomari, and Y. Hadjadj-Aoul, "A consistent QoS routing strategy for video streaming services in SDN networks," *Int. J. Commun. Syst.*, vol. 33, no. 10, p. e4177, Jul. 2020.
[209] R. S. Kalan and M. Sayit, "SDN assisted codec, path and quality selection for HTTP adaptive streaming," *IEEE Access*, vol. 9, pp. 129917–129932, 2021.
[210] M. O. Elbasheer, A. Aldegheishem, N. Alrajeh, and J. Lloret, "Video streaming adaptive QoS routing with resource reservation (VQoSRR) model for SDN networks," *Electronics*, vol. 11, no. 8, p. 1252, Apr. 2022.
[211] R. Farahani, M. Shojafar, C. Timmerer, F. Tashtarian, M. Ghanbari, and H. Hellwagner, "ARARAT: A collaborative edge-assisted framework for HTTP adaptive video streaming," *IEEE Trans. Netw. Service Manage.*, vol. 20, no. 1, pp. 625–643, Mar. 2023.
[212] P. H. Thinh, N. T. Dat, P. N. Nam, N. H. Thanh, H. M. Nguyen, and T. T. Huong, "An efficient QoE-aware HTTP adaptive streaming over software defined networking," *Mobile Netw. Appl.*, vol. 25, no. 5, pp. 2024–2036, Oct. 2020.
[213] M.-H. Wang, T.-S. Hsieh, Y.-Y. Tseng, and P.-W. Chi, "A software-defined networking-driven reliable transmission architecture for enhancing real-time video streaming quality," *IEEE MultimediaMag.*, vol. 31, no. 1, pp. 54–64, Jan. 2024.
[214] R. S. Kalan, "Improving quality of HTTP adaptive streaming with server and network-assisted DASH," in *Proc. 17th Int. Conf. Netw. Service Manage. (CNSM)*, Oct. 2021, pp. 244–248.
[215] R. Farahani, F. Tashtarian, A. Erfanian, C. Timmerer, M. Ghanbari, and H. Hellwagner, "ES-HAS: An edge- and SDN-assisted framework for HTTP adaptive video streaming," in *Proc. 31st ACM Workshop Netw. Operating Syst. Support Digit. Audio Video*, Jul. 2021, pp. 50–57.



REZA KALAN received the B.S. degree in computer engineering from IAU University, Iran, in 2003, the M.S. degree in computer engineering from EMU University, Cyprus, in 2013, and the Ph.D. degree in information technology from Ege University, Türkiye, in 2020. From 2019 to 2024, he was a Researcher with Digiturk Research and Development. Since 2024, he has been an Assistant Professor with the Computer Engineering Department, Istinye University. He is the author of three books. His research interests include computer networks, multimedia systems, software-defined networks, and mobile communication. He focuses mainly on resilient and secure networking, with interests in future internet and cellular communication. He has served on the TPC of top conferences, such as NoF, NetSoft, and NOMS.

ISMAIL DULGER received the M.Sc. degree in computational chemistry from Gebze Technical University. He is currently pursuing the Ph.D. degree with Ege University, further advancing his expertise in video streaming technologies. He is also a Research and Development Coordination Specialist at Digiturk, where he focuses on research and development in the field of video streaming technologies.